University of Alberta

Speaker Information Enhancement

by

Fangxin Chen ©

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Doctor of Philosophy

in

Speech Production and Perception

Department of Linguistics

Edmonton, Alberta

Fall 1997

National Library of Canada
Acquisitions and Bibliographic Services
395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale du Canada
Acquisitions et services bibliographiques
395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-22964-5

Canada

University of Alberta

Library Release Form

Name of Author: Fangxin Chen

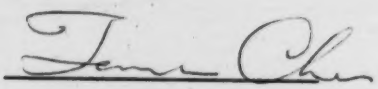
Title of Thesis: Speaker Information Enhancement

Degree: Doctor of Philosophy

Year this Degree Granted: 1997

Permission is hereby granted to the University of Alberta Library to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly, or scientific research purposes only.

The author reserves all other publication and other rights in association with the copyright in the thesis, and except as hereinbefore provided, neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatever without the author's prior written permission.



202, 10615-40Av
Edmonton, Alberta
Canada T6J 2W3

July 28, 1997

UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled SPEAKER INFORMATION ENHANCEMENT submitted by Fangxin Chen in partial fulfilment of the requirements for the degree of DOCTOR OF PHILOSOPHY in SPEECH PRODUCTION AND PERCEPTION.

John T. Hogan, supervisor

Bernard L. Rochet, committee member

Aleksandar Kostov, committee member

Xiaobo Li, committee member

Douglas O'Shaughnessy, external examiner

July 7, 1997
(date)

This thesis is dedicated to

my family

Abstract

This study proposes a new method for speaker-information enhancement in computer speaker recognition. The basic approach is to measure the distribution pattern of speaker information in the parametric domain with the training speech data, and then apply corresponding weighting strategy in the testing phase to enhance those speaker-information rich elements. This approach is based on the assumption that a speech signal contains both phonetic (linguistic) and speaker information. Though interrelated, these two kinds of information have their own distinctive representations in the acoustic and parametric domain. For speaker recognition purposes, only the speaker-information component in the speech signal should be maximally enhanced.

This study first reviews the anatomical, psychological and social nature of a speaker's voice quality and its manifestation in the acoustic domain. The phonetic environment and the interrelationship between phonetic and speaker information are also discussed. Since the cepstral representation of a speech signal is widely used in present speech and speaker recognition applications, the new approach for speaker-information enhancement is tested in the cepstral domain.

The experiment consists of two parts. The first part investigates speaker-information distribution in the mel frequency cepstrum coefficients (MFCC). Two statistical methods, *the inter-distribution distance measurement* and *the speaker identification error rate measurement*, are used independently for estimating the amount of speaker information coded in each MFCC coefficient. The second part

of the experiment compares three different cepstral weighting (or liftering) approaches for speaker-information enhancement. The first weighting is a traditional speech weighting method; the other two weightings are based on either the general or individual speakers' distribution patterns of speaker information. The experimental results indicate that the weightings based on speaker information provide better speaker recognition performance.

Though the proposed new approach for speaker information enhancement was tested in the cepstral domain, the same principle for maximally enhancing the speaker-related variability should also be applicable to other speech parameters. In this respect, the present study provides a methodology, rather than only a specific technique, for speaker information enhancement. This new approach also has its implication for speech recognition research.

Acknowledgments

When I have finished this thesis, I feel I owe a debt of gratitude to many people. First of all, I would like to express my sincere thanks to John T. Hogan, my thesis supervisor, who has shown constant support and encouragement for my thesis work. His guidance has been a valuable source of inspiration for me.

I am very grateful to my thesis committee members. I was blessed with a support team with expertise in different disciplines. The quality of my thesis has significantly benefited from Bernard L. Rochet's linguistic insights, Aleksandar Kostov's engineering perspective, and Xiaobo Li's knowledge in pattern recognition.

I am also very grateful to my external examiner Douglas O'Shaughnessy, an eminent scholar in the field of speech technology. I feel fortunate to have had him reviewing my thesis and attending my thesis defense. His critical comments were particularly helpful for my thesis revisions.

My gratitude also goes to Anton J. Rozsypal, who introduced me into this exciting field of computer speech technology, and to Bruce Millar, who gave me the wonderful opportunity to work in a speech research project at the Australian National University.

Finally, I would like to express my appreciation to all the people at the Department of Linguistics who have shown their support, encouragement and friendship in my academic endeavor.

Table of Contents

Chapter 1 Introduction 1

Chapter 2 Speaker Information in the Acoustic Domain 4

2.1 Physical Markers 6

2.2 Psychological Markers 17

2.3 Social Markers 19

2.4 Speaker Vs. Phonetic Information 21

2.5 Intra-Speaker Variation 24

Chapter 3 Speaker Information in the Phonetic Domain 29

3.1 Vowels 29

3.2 Fricatives 31

3.3 Nasal Consonants 32

3.4 Stops 32

3.5 Quantitative Studies 34

Chapter 4 Speaker-Information Extraction 35

4.1 Noise-Reduction Approach 35

4.2 Auditory Approach 37

4.3 Phonetic Approach 38

4.4 Speaker-Information Enhancement Approach 39

Chapter 5 Cepstral Representation of the Speech Signal 42

5.1 Parametrization 42

5.2 Spectral Representation of the Speech Signal 42

5.3 Homomorphic Filtering 47

Chapter 6 Speaker-Information Distribution in the Cepstral Domain 50
6.1 Experimental Design 50
6.2 Speech Database 51
6.3 Data Pre-Processing 54
6.4 Prarmetric Representation of the Speech Signal 55
6.5 Speaker Modelling 57
6.6 VQ Distortion Score Measure 60
6.7 Inter-Distribution Distance Measurement 61
6.8 Speaker Identification Error Rate Measurement 67
6.9 Phonetic-Information Distribution 72
6.10 Discussion 75

Chapter 7 Speaker-Information Enhancememt 79
7.1 Experimental Design 79
7.2 Experimental Results 86
7.3 Discussion 86

Chapter 8 Summery and Conclusion 92

Bibliography 95

Appendices 108
Appendix I-A: Individual speakers' NIDD score distributions in the cepstral coefficients. 108
Appendix I-B: Individual speakers' NIER score distributions in the cepstral coefficients. 108
Appendix II-A: Individual speakers' NIDD score distributions in the cepstral coefficients [female speaker group]. 109

Appendix II-B: Individual speakers' NIDD score distributions in the cepstral coefficients [male speaker group]. 110
Appendix III-A: Individual speakers' NIER score distributions in the cepstral coefficients [female speaker group]. 111
Appendix III-B: Individual speakers' NIER score distributions in the cepstral coefficients [male speaker group]. 112
Appendix IV: Individual speakers' NRER score distributions in the cepstral coefficients. 113
Appendix V-A: Individual speaker's NRER score distribution in the cepstral coefficients [female speaker group]. 114
Appendix V-B: Individual Speaker's NRER Score Distribution in the cepstral coefficients [male speaker group]. 115

List of Tables

Table 2-1 Individual speakers' average F_0s and STDs for 16 repetitions of the word "zero" from the TI-46 speech data. 8

Table 2-2 Formant frequencies of [i] for four female speakers and four male speakers from the TI-46 speech data. 14

Table 7-1 Cepstral coefficient weighting for individual speakers based on the NIER score. 84

Table 7-2 Cepstral coefficient weighting for individual speakers based on the NIDD score. 85

Table 7-3 Speaker identification error rates for different weighting strategies. 86

List of Figures

2-1	Male and female Fo differences.	9
2-2	DFT and LPC spectra of the vowel [i] spoken by four female and four male speakers.	13
2-3a	The spectrogram of a normal female speaker's speech "x-ray".	17
2-3b	The spectrogram of a dysarthric female speaker's speech "x-ray".	17
2-4	Variability of female and male speakers' F_0.	20
2-5	Long-time spectra of six utterances of the same word " zero" spoken by a female speaker in the TI-46 data.	26
3-1	The average female/male formant ratios of F1, F2, and F3 from American English and Swedish data. (after Fant, 1973).	30
5-1	(a) Waveform for a short interval of the vowel [i]; (b) the DFT of the speech signal.	44
5-2	Spectral representation of a short interval of the vowel [i] using LPC Analysis.	45
5-3	DFT and LPC spectral representations of a short interval of the vowel [i].	46
6-1	Phonetic composition of the TI-20 vocabulary.	53
6-2	Block diagram of MFCC processing.	55
6-3	Mel-Scale Filter Bank.	56
6-4	Block diagram of VQ codebook training.	59

6-5 Variance distribution in the cepstral coefficients. 61

6-6 A case of intra-speaker and inter-speaker VQ distortion score distribution pattern. 62

6-7 Average NIDD score distribution in the MFCC over all the speakers. 65

6-8 Average NIDD score distribution in the MFCC over the female speaker group. 66

6-9 Average NIDD score distribution in the MFCC over the male speaker group. 66

6-10 Block diagram of the speaker identification error rate measurement. 69

6-11 Average NIER score distribution in the MFCC over all the speakers. 70

6-12 Average NIER score distribution in the MFCC over the female speaker group. 71

6-13 Average NIER score distribution in the MFCC over the male speaker group. 71

6-14 Block diagram of the speech recognition error rate measurement. 73

6-15 The average NRER score distribution in the MFCC over all the speakers. 74

6-16 The average NRER score distribution in the MFCC over the female speaker group. 75

6-17 The average NRER score distributions in the MFCC over the male speaker group. 75

6-18 Speaker F1's phonetic- and speaker-information distributions in the MFCC. 77

7-1 Raised sine weighting function. 81

7-2 Weighting function based on the average NIER score distribution in the MFCC coefficients. 82

7-3 Weighting function based on the average NIDD score distribution in the cepstral coefficients. 83

7-4 Comparison of speaker identification performance between the baseline and Weighting Function A. 87

7-5 Comparison of speaker identification performances among the baseline, Weightings Function B1 and B2. 88

7-6 Comparison of the speaker identification performances among the baseline, Weighting Functions C1 and C2. 90

CHAPTER 1

INTRODUCTION

The goal of computer speaker recognition is to determine "Who is speaking?" (speaker identification) or "Is the speaker the claimed client?" (speaker verification). This is an area of artificial intelligence where machine performance can exceed human performance (O'Shaughnessy, 1987). Because of the potential applications in high-security data or area access, bank transactions, telephone services etc., there has been in recent years an increased momentum in speaker recognition research. From an engineering perspective, speaker recognition shares quite similar techniques with speech recognition, such as digital signal processing, parametrization, statistical modelling and pattern recognition. From a linguistic point of view, however, these two tasks differ in nature. The task of speech recognition is to search for phonetic information, i.e., the invariant acoustic cues in the speech signal. Speaker variability is considered a confounding factor which needs be suppressed in the feature extraction and pattern recognition processes. For speaker recognition, on the contrary, speaker variability provides all the information needed for speaker identification and verification, while phonetic information becomes an irrelevant or even confounding factor. How to effectively extract speaker information (or speaker variability) from the speech signal, then, is the fundamental question for speaker recognition research. The literature has reported various methods for enhancing speaker recognition performance, such as selecting the optimal parametric representation of the speech signal and using sophisticated statistical methods for speaker modelling and distance measurement (Rosenberg & Soong, 1992; Furui, 1994). However, the fundamental question of how speaker idiosyncrasy is coded in the speech parameters and the effective method to extract it have not yet been sufficiently exploited for the benefit of improving speaker recognition performance.

The present study proposes a new approach, originally inspired by Furui (1994), for speaker-information enhancement. In his overview of speaker recognition technology, Furui suggested pursuing " a method for extracting and representing the speaker characteristics that are commonly included in all the phonemes irrespective of the speech text" (p. 8).

The new approach is based on the assumption that speech contains both speaker- and phonetic-information components. Though interrelated, these two kinds of information have their own characteristic distribution patterns in the acoustic domain. For speaker recognition, only the speaker-information component in the speech signal should be enhanced.

The new approach for speaker-information enhancement can be briefly described as using the training speech data to measure the distribution of speaker information in the parametric domain, and then applying the optimal weighting strategy in the testing phase to enhance only the speaker-information rich elements. Since the cepstrum is a widely used parametric representation of the speech signal in both speech and speaker recognition systems, the present study tests the new approach in the cepstral domain. Efforts are focused on the measurement of speaker information in the mel frequency cepstrum coefficients (MFCC), and the optimal weighting strategy for speaker-information enhancement. The study is divided into the following eight chapters:

Chapter 1 is the introduction of the study.

Chapter 2 surveys the anatomical, psychological and social origins of a speaker's voice quality and their representations in the acoustic domain.

Chapter 3 investigates the interrelationship between speaker information and its phonetic environment.

Chapter 4 is a review of the methods for speaker-information extraction, followed by the proposal of a new approach for speaker-information enhancement.

Chapter 5 discusses the technical aspects of the cepstral representation of the speech signal.

Chapter 6 presents the experiments for measurement of speaker information in the cepstral domain. For comparison, the distribution pattern of phonetic information is also investigated.

Chapter 7 compares different weighting strategies for an optimal speaker-information enhancement.

Chapter 8 presents the summary and conclusions.

CHAPTER 2

SPEAKER INFORMATION IN THE ACOUSTIC DOMAIN

A speech signal carries both phonetic and speaker information. Phonetic information is related to three distinct linguistic functions(Trubetzkoy, 1969).

- the meaning-differentiating (distinctive) function, which distinguishes the individual units of meaning.
- the culminative function, which indicates the important linguistic units contained in a particular utterance.
- the delimitative function, which signals the boundaries between the linguistic units.

Speaker information refers to a speaker's *voice quality* defined by Laver (1980) as the characteristic auditory colouring of an individual speaker's voice. Acoustically, speaker information is reflected in an individual speaker's idiosyncratic spectral representation of speech. Speaker information permits the hearer to identify individual speakers.

Compared to phonetic information, speaker information is usually not under a speaker's conscious control. Speaker information reflects mostly the invariant aspects of a speaker's anatomical structure and the habituated nature of vocal settings. The function it plays in speech is *informative* rather than *communicative*. The distinction between *communicative* and *informative* was defined by Lyons (1977) as : a signal is *informative* if (regardless of the intentions of the sender) it makes the receiver aware of something of which he was not previously aware; on the other hand, a signal is *communicative* if it is intended by the sender to make the receiver aware of something of which he was not previously aware.

Besides its informative function, speaker information also plays a "channel" function in speech communication. A well-known example is the "cocktail party effect" (Ainsworth, 1976). Without a speaker cue, it would be difficult to follow a selected conversation when the surrounding conversations are of greater loudness.

Sapir (1927) distinguished different levels in the structure of the voice: voice proper, intonation, rhythm, continuity and speech rate. He suggested that each of these has individual and social dimensions that interact to create various voice patterns.

O'Shaughnessy (1987) attributed speaker variations to three sources: differences in vocal cords and vocal tract shape, differences in speaking style and differences in what the speaker chooses to say.

Rosenberg and Soong (1992) classified speaker information into two categories: the *physical* and *behavioural* information of the speaker. The *physical information* refers to a speaker's speech-related anatomical structures such as the oral cavity, vocal folds, velum and nasal cavity; the teeth and jaw configuration; the respiratory volume; the dimensions of lips and the geometry of laryngeal structures. Behavioural information can be low-level or high-level. Low-level behavioural information is associated with vocal settings such as tongue position, pitch contours, rhythm, etc. . High-level behavioural information takes the form of the characteristic choice of words, phrases and other aspects of speech style.

Laver and Trudgill (1979), in a more philosophical approach, characterised speaker information in a speech signal as three types of markers: *physical markers*, *psychological markers*, and *social markers*. *Physical markers* are similar to Rosenberg and Soong's definition of physical information, which refers to a speaker's physical characteristics, such as vocal anatomy, age, sex, physique and state of health; *psychological markers* refer to those marking a speaker's psychological characteristics of personality, and affective and attitudinal status; *social markers* refer to those marking a speaker's social characteristics such as regional affiliation, social and educational status, occupation and social role. The same voice phenomenon can be attributed to different types of markers. For example, a speaker's particular vocal setting can be either psychological or social. The following discussion of speaker information will adopt Laver and Trudgill's typology.

2.1 Physical Markers

Speech production is a complex co-ordination of muscles and speech organs. Speech-related muscle systems can be classified into following categories (Laver and Trudgill, 1979):

- the *Respiratory System* which co-ordinates the lungs' inspiration and expiration.
- the *Phonatory System* which controls the larynx during phonation.
- the *Pharyngeal System* which controls articulatory activity at the posterior end of the vocal tract.
- the *Velopharyngeal System* which controls the production of nasality.
- the *Lingual System* which is responsible for the oral articulations.
- the *Labial System* which controls the actions of the lip structures.
- the *Mandibular System* which controls movements of the jaw.

The speech organs consist of *the subglottal respiratory system*, *the larynx* and *the superlaryngeal vocal tract* (Lieberman and Blumstein, 1988). They are the apparatus for generating speech and also the main sources for individual variability. The description of physical markers will be focused on these three aspects.

2.1.1 The Subglottal Respiratory System

The subglottal respiratory system (SRS) consists of the lungs and associated respiratory musculature. The function of the SRS in speech is to provide an airstream for phonation. Speech sounds are the result of manipulation of air stream from the SRS. One speaker difference in the SRS is the *vital capacity*, which is measured by taking a maximum inspiration, then measuring a maximum expiration. *Vital capacity* is related to a speaker's sex, size, and breathing habits. Another speaker variability in the SRS is the individual pattern in controlling the subglottal air pressure. Intensity of voicing will increase as a function of a value between the 3rd and 4th power of the subglottal air pressure, which means that a

small change in subglottal air pressure makes a large intensity difference. Different intensity patterns of speech by different speakers can be observed perceptually or acoustically. Fant (1973) found that an increase in subglottal air pressure did not cause a uniform increase of the intensity level over all the frequency ranges. In general, a greater intensity increase was observed in the higher frequency range than in the lower frequency range.

The relationship between subglottal air pressure and the fundamental frequency (F_0) was addressed by many experimental and modelling studies (Strik and Boves, 1992). The F_0 to subglottal air pressure ratio was estimated in values between 5 - 15 Hz/cm H_2O.

Perkell et al. (1994) reported sex differences in subglottal air pressure and maximum airflow declination rate. As for age differences in speech respiration, an investigation by Hoit and Hixon (1987) revealed that ageing men generally used larger lung volume excursions per breath group, and expended larger lung volumes per syllable. In addition, fewer syllables per breath group and the initiation of breath groups from higher rib cage volumes were observed. In speech pathology, it was reported that a creaky voice (vocal fry) was associated with low subglottal air pressure (Laver, 1980).

2.1.2 The Larynx

The larynx consists of the thyroid, cricoid and arytenoid cartilages and the vocal folds, which are shelf-like elastic protuberances of tendon, muscles and mucous membranes. According to the myoelastic aerodynamic theory of phonation proposed by Van den Berg(1958), the vibration of vocal folds is caused by two forces: the myoelastic tensions acting on the vocal folds, and the aerodynamic Bernoulli effect. With the glottis closed by muscular tensions, sub-glottal air pressure builds up. When the pressure reaches certain level, it blows open the vocal folds. The air stream passes the narrow glottis, and the increase of velocity causes the sudden pressure drop. Then, due to the Bernoulli effect, as well as the myoelastic tensions, the vocal folds are pulled together. The closure of the

glottis causes the air pressure to build up again, leading to the continued vibration of vocal folds. The frequency of vocal-fold vibration is determined by the elasticity, tension, and mass of the vocal folds. Long and thick vocal folds vibrate at lower F_0. Since males usually have longer and thicker vocal folds than females do, males usually have a lower natural F_0 than females. To examine F_0 differences between male and female speakers, the F_0 s of 16 repetitions of the word "zero" spoken by eight male and eight female speakers, respectively, were extracted using the Cspeech (Milenkovic and Read, 1992). The average F_0 and the standard deviation (STD) for each speaker are presented in Table 2-1 and also plotted in Figure 2-1. The Speech data used is TI-46, which will be described in detail in Chapter 6.

Table 2-1: Individual speakers' average F_0s and STDs for 16 repetitions of the word "zero" from the TI-46 speech data

Speaker	Fo (Hz)	STD (Hz)	Speaker	Fo (Hz)	STD (Hz)
F1	179	19.98	M1	109	25.36
F2	237	30.63	M2	123	12.6
F3	230	21.59	M3	132	20.88
F4	156	20.44	M4	129	13.4
F5	188	15.68	M5	102	19.42
F6	198	33.46	M6	120	17.11
F7	220	15.88	M7	130	9.31
F8	199	35.97	M8	132	32.14
Average	200.88	24.2	Average	122.13	18.78

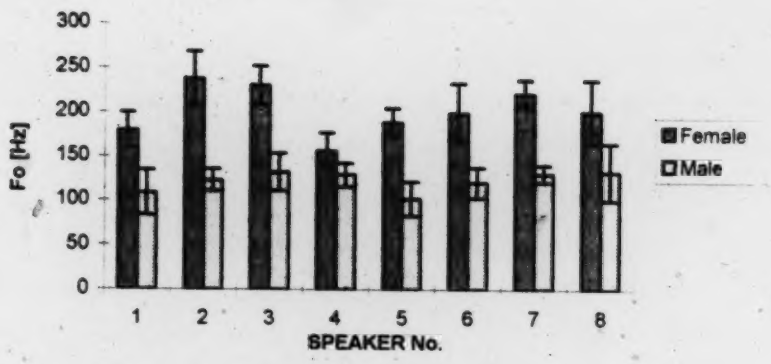


Figure 2-1 Male and female Fo differences.

It can be observed from Figure 2-1 that all the eight female speakers have a higher F_0 than the male speakers. The average F_0 over all the male speakers is 122 Hz, while the average F_0 over all the female speakers is 201 Hz. In each sex group, there are also individual variations. The average F_0 range for the male speakers is between 102 Hz to 132 Hz, and the average F_0 range for the female speakers is between 156 Hz to 237 Hz. This result is close to the F_0 value reported by Spencer's perceptual experiment (1988). In that perceptual experiment, male-to-female transsexuals' voices were presented to the listeners, who were asked to judge whether the voices were male or female voices. The result was that all speakers with F_0 below 160 Hz were identified as male, and all those speakers with F_0 above 160 Hz were identified as female.

In addition to sex information, F_0 also carries age information. Children usually have higher natural F_0 than adults because children's vocal folds are not well developed, i.e. , children's focal folds are much shorter, thinner and more elastic than adults'. For male speakers, there is a general lowering trend of F_0 from infancy through middle age, then a reversal in the trend whereby the F_0 rises slightly with advancing age. With females, however, there is controversy over whether F_0 changes with advancing age. Some experiments with aged females found no significant F_0 changes (Mcglone and Hollien, 1963; Biever and Bless,

1989). The explanation is that anatomical changes in the female larynx are not as extensive as those for men. Therefore, the degenerative changes may not have as great an effect on females' laryngeal structures as on males' in their advanced ages. On the other hand, other experiments (Honjo and Isshile, 1980; Bussel et al. , 1995) found that aged females' F_0 was significantly lower than young females'. Bussel et al.'s experiment used archival data, recordings made in 1945 of a group of Australian females, and recordings made in 1993 of the same group of females. The result showed significant lowering of F_0 with aging. However, using recording data from about 50 years ago could be problematic if some technical aspects were not strictly controlled.

The aging factor is also reflected in F_0 variability. Endres et al. (1971) showed that the individual F_0 distribution curves became narrower with increasing age because of the decreasing ability in controlling the laryngeal muscles.

The F_0 perturbation can also be informative about a speaker's vocal folds' anatomical characteristics. Jitter is the cycle-to-cycle variation of the glottal period, and shimmer is the cycle-to-cycle variation of the peak amplitude. The aperiodic vocal fold vibration gives a voice the quality of harshness. Such vocal fold vibration is caused by excessive laryngeal tension. The jitter and shimmer patterns for individual speakers are quite different.

The anatomy of the larynx affects the glottal source waveform characteristics. Craner and Schroeter (1995) showed that in the speech spectrum, there exists a *dc* component during the "closed glottis" interval of vowels produced by both male and female speakers at a normal loudness level. This *dc* component is caused by glottal leakage. A moderate leakage may give rise to appreciable source-tract interaction. In the time domain, such leakage manifests itself as a ripple in the glottal flow waveform just after closure. In the frequency domain, the spectrum of the flow through a glottis with a leak is characterized by anti-resonances. Hanson (1997) found that incomplete glottal closure causes

- an increase in the bandwidth of the first (and possibly the second) formant.
- an increase in the tilt of the glottal spectrum at high frequencies.

- an emergence of a turbulence noise source in the vicinity of the glottis that may be comparable in amplitude (at high frequencies) to the spectrum amplitude of the periodic source.

A more open glottal configuration results in relatively greater low frequency and weaker high-frequency components, compared to a more adducted glottal configuration. The more open glottal configuration also leads to a greater source of aspiration noise and larger bandwidth of the natural frequencies of the vocal tract, particularly the first formant. Female speakers are found more likely than male speakers to have incomplete closure of the focal folds. The degree of incomplete closure of the vocal folds is also speaker-dependent.

2.1.3 The Vocal Tract

The vocal tract includes the pharyngeal cavity, the oral cavity and the nasal cavity. According to the source-filter theory (Fant, 1960), the vocal tract acts as an acoustic resonator. If *G (f)* represents the glottal source spectrum, *H (f)* represents the gain factor of the filter function at the frequency of *f*, then the speech spectrum *X(f)* can be estimated as

$$X(f) = G(f) * H(f) . \qquad (2.1)$$

The spectral peaks of X (f) are called "formants" or "poles". They correspond to the resonances of the vocal cavity, which can be approximated by a tube open at one end and closed at the other. For a vocal tract with a fairly uniform cross-section, the resonances or the formant frequencies can be estimated as

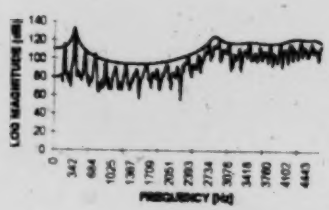
$$F_i = \frac{(2i-1)c}{4L} . \qquad (2.2)$$

where F_i is the *i*th resonance (or formant frequency); *c* is the velocity of sound in the air and *L* is the length of the vocal tract. For the schwa sound [ə] produced

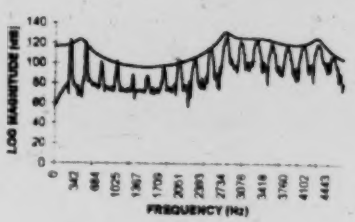
by a male with vocal tract length of 17 cm, F_1 will be around 500 Hz, F_2 around 1500 Hz, F_3 around 2500 Hz, and so forth. The average frequencies of the vowel formants are inversely proportional to the length of the vocal tract. Fant (1973) found that:

> The percentage of relative increase in formant frequencies associated with the removal of one-half centimetre of the pharynx of the vowel [i] is 3.5% in F_1, 4.7% in F_2, and 0.5% in F_3. Similarly, a removal of a one-half centimetre section of the frontal mouth cavity of [i] results in a 1.3% increase in F_1, a 0.2% increase in F_2, and 6.1% increase in F_3. (p.88)

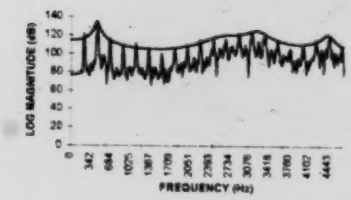
Since different speakers have vocal tracts of different lengths and shapes, the natural resonant frequencies or formants will be different among speakers. The average female vocal tract is about 15% shorter than the male's vocal tract; and consequently, females usually have higher average formant frequencies than males. Anatomical studies have also found that males have relatively greater pharynx length and more pronounced laryngeal cavities than females, which also contributes to the formant differences between sexes. We analysed a segment of speech spoken by four female speakers and four male speakers respectively from the TI-46 speech data. Phonetically, this segment is the steady state portion of the vowel [i] with duration of 100 ms. We did both Discrete Fourier Transform (DFT) and Linear Predictive coding (LPC) analyses, and the DFT and LPC generated spectra are plotted in the graphs of Figure 2-2. The measured formant frequencies of [i] for each speaker are presented in Table 2-2.



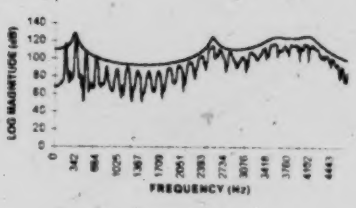
[Female 1]



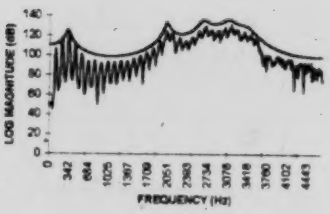
[Female 2]



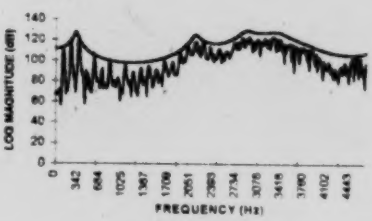
[Female 3]



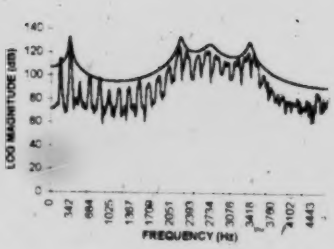
[Female 4]



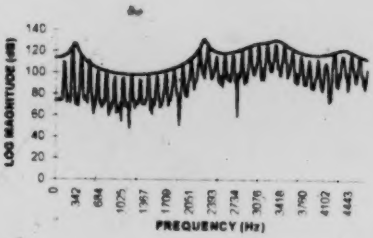
[Male1]



[Male 2]



[Male 3]



[Male 4]

Figure 2-2 DFT and LPC spectra of the vowel [i] spoken by four female and four male speakers. In each graph, the highly varying spectrum is from the DFT spectral analysis, and the smooth spectrum from the LPC spectral analysis. The spectral peaks are the formants.

Table 2-2: Formant frequencies of [i] for four female speakers and four male speakers from the TI-46 speech data

Speaker	F_1 (Hz)	F_2 (Hz)	F_3 (Hz)	F_4 (Hz)
Female 1	366	2832	3491	4285
Female 2	427	2771	3137	4297
Female 3	452	2865	3210	4407
Female 4	330	2527	3528	4089
Average [female]	**394**	**2749**	**3342**	**4270**
Male 1	305	2026	2673	3101
Male 2	317	2112	2905	3320
Male 3	330	2209	2698	3369
Male 4	281	2234	3000	3333
Average [male]	**308**	**2145**	**2819**	**3281**

Figure 2-2 and Table 2-2 indicate the formant difference between the sex groups. The average formant values of the female group are significantly higher than those of the male group. Individual differences also exist within each sex group.

If both the oral and nasal cavities are involved in articulation, anti-resonances or "zeros" are introduced into the spectrum because the side-branching chamber acts as a filter, and selectively absorbs energy from the main tube at frequencies which are dependent, in part, on the side-chamber's own resonant frequencies. In speech production, either the nasal cavity or the oral cavity can act as a side-chamber relative to the other (Laver, 1980). Provided that the exit of the cavity is smaller than the entrance, whichever cavity has the smaller exit becomes the side-chamber. Spectrally, the anti-resonances or "zeros" cause the formant frequencies to be dampened. Since a speaker's nasal cavity, compared with the oral cavity, is not subject much to change during articulation, the resonances or anti-resonances of the nasal cavity can provide quite reliable speaker information.

2.1.4 Pathological Status

A speaker's pathological status might also be determined from the speech signal. Partial vibration of the vocal cords creates a creaky or laryngealized voice. Acoustically, a creaky voice is manifested by a narrowing of the glottal pulse in its duration and a lowering of F_0 and its amplitude. Comparing creaky voices of both sexes, Klatt and Klatt (1990) found no significant F_0 range difference, and suggested that in the creaky mode, F_0 is not affected by larynx variations. When the vocal folds are irritated and swollen, the voice becomes hoarse. Incomplete glottal closure during the "closed" phase of the phonation gives a breathy voice quality. Breathy glottal source signals obtained through inverse filtering typically show more symmetrical opening and closing phases with little or no complete closed phase. With increasing breathiness in speech, the F_0 tends to be lowered and the relative F_0 amplitude tends to be increased. As well, the first formant bandwidth can also be increased. With the air passing the narrow glottal chink during the "closed" phase of phonation, noise is generated. This reduces the amplitude of the higher frequency harmonics, and the upper portion of the spectrum becomes dominated by a dense aspiration noise. In a breathy voice, more high frequency energy is evident than in a normal voice (Klatt and Klatt, 1990; Hillernbrand et al. , 1994).

Apart from pathologies specific to speech, other health problems of a speaker might also be reflected in his/her speech. Speakers suffering from Parkinson's disease, for example, have reduced stress in speech. Stress reduction may result from respiratory/phonatory impairments, such that pitch and loudness, which cue syllabic stress, are diminished. For hearing-impaired speakers, Subtelny et al. (1989) found that their tongue tended to retract for front vowels and to move frontward for back vowels. For high vowels, most of the hearing-impaired speakers had an elevated hyoid, an unusually large vertical dimension between hyoid and laryngeal sinus, and a retracted tongue root associated with a marked retraction or deflection of the epiglottis toward the pharyngeal wall. The characteristic voice of persons with Down Syndrome is harsh, raucous and low-

pitched. The spectrum shows the emphasizing of the F_0 for all vowels, and the range of the spectrum is narrowed. The irregular vibration of the vocal folds causes the irregular breaks in the F_1. The F_2 is often very weak, and the higher formants are invisible (Novak, 1971). Dysarthric speakers usually show in their articulation (1) anterior lingual place inaccuracy, (2) reduced precision of fricative and affricate features, (3) an inability to achieve extreme positions in vowel articulatory space, (4) nasalization of vowel production (Andrews et al. , 1977; Yorkston et al. , 1989).

The nasalization of dysarthric speech is possibly caused by the inability of the patient's velopharyngeal musculature to control the velum for closing the velopharyngeal port, resulting in the air leaking through the nasal cavity. Weismer et al. (1995) observed that in most motor disorders, articulatory gestures are typically slow; the articulatory gestures are often under-scaled; and there are significant intra- and inter-speaker variations. Figure 2-3 illustrates the spectrograms of a normal speech and a speech manifesting a motor disability. Figure 2-3a is the spectrogram of the word " x-ray " spoken by a normal female speaker. Figure 2-3b is the spectrogram of the same word spoken by a female speaker with dysarthria. Comparison of the two spectrograms clearly shows that the dysarthric speech has much longer utterance duration (about double the normal utterance duration). The fricative [s] spectrogram is observable in the normal speech, but nearly disappears in the dysarthric speech. In addition, the dynamic transition in the diphthong [ei] is observable in the normal speech spectrogram, but is flattened in the dysarthric speech spectrogram.

Figure 2-3.a The spectrogram of a normal female speaker's speech "x-ray".

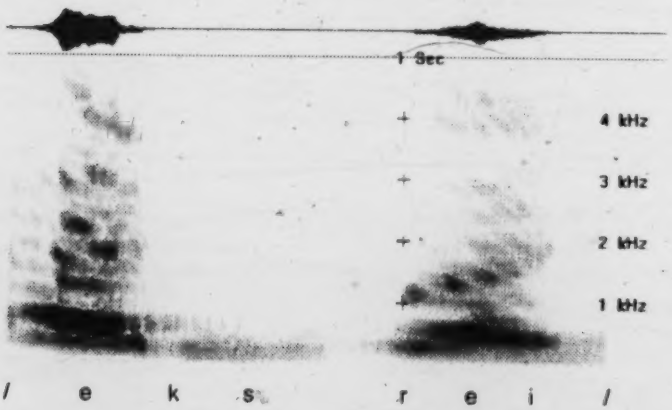


Figure 2-3.b The spectrogram of a dysarthric female speaker's speech "x-ray".

2.2 Psychological Markers

A speaker's psychological status or personality may also be reflected in the speech signal. Early studies showed that breathy voices might be indicative of introversion, neurotic tendency and anxiety (Diehl et al. , 1959; Moor, 1939). Siegman and Pope (1965) found extroversion to be associated with shorter

latency, fewer 'filled' brief pauses and fewer silent pauses. F_0 was reported to have a strong positive correlation with a speaker's personality traits such as sociability, dominance and aggressiveness (Scherer & Giles, 1979). Competent and dominant speakers usually have a higher F_0 than non-task-oriented, submissive speakers. In addition, F_0 can be an indicator of the stress level of the speaker. Scherer showed that stress-induction led to a significant rise of speaker's F_0. Laver (1975) found that for tense voices caused by high muscular tension when a speaker was in an emotional state, the energy concentration was between 500 and 1000 Hz, compared with a lax voice, which had energy concentration below 500 Hz. A speaker's psychological stress can also affect the magnitude of shimmer and jitter (Inbar and Eden, 1976). According to Williams' research (1972), a speaker's emotion has its acoustic correlation in the speech signal. In a neutral emotion state, vowels tend to have well-defined formant structures, with little noise or irregularities, either between the formants or in the high frequency regions. Consonants are uttered in an imprecise manner, particularly when in unstressed syllables. The F_0 contour, as a function of time, is characterized by smooth, slow and continuous changes. The F_0 changes occur with syllable stress or semantic emphasis. In an emotional state, both the spectral and F_0 patterns are affected. For example, if the speaker is angry, high F_0 is observed throughout the breath group. F_0 increase is, on the average, at least half an octave above that during the neutral state. F_0 variation is also considerably larger. Some syllables are produced with significantly increased intensity. The vowels in these syllables tend to be articulated with a more open vocal tract. The F_1 is usually weakened in intensity and increased in frequency, while the consonants are generated with a more clearly defined closure. A more recent study (Laukkanen et al. , 1996) suggested that stress or emotion brought about simultaneous changes in F_0, SPL, subglottal pressure and the glottal airflow waveform.

2.3 Social Markers

A speaker's social status, such as education, occupation, regional affiliation and social role, is also marked in the speech signal. Social and regional dialects are the obvious examples. In his investigation of Norwich English, Trudgill (1974) found that the working-class speakers tended to use a creaky voice, a high pitch range, a wide loudness range, a lowered tongue position, a raised larynx position, a particular type of nasality and a relatively high degree of muscular tension, which distinguish them from the middle-class speakers.

Cultural differences are also reflected in a speaker's speech pattern. Laver (1975) reported a very low pitch range in American males. Scherer (1979) found a significant difference in the mean F_0 for his American and German speakers. For his 28 American subjects, the mean F_0 was 128 Hz and for his 29 German subjects, the mean F_0 was 161 Hz. The cultural markers in speech reflect a speaker's learned behaviour and can be attributed to historical tradition, cultural stereotypes, national character or even the influence of television and movie stars. Moreover, people with certain occupations could have unique speech patterns. For example, Kuwabara et al. (1983) analysed the speech recordings of several announcers, finding that their speech could be characterized by the dynamic characteristics of the F_0 and formant frequencies. Compared with people of other occupations, announcers' voices also had a higher energy level in the 3-4 kHz frequency band.

Some of the sex differences in speech can also be culture markers. For example, females are more likely than males to have incomplete closure of the vocal folds in their speech, causing an airflow bypass even during the closed phase of the glottal vibratory cycle and creating a breathy voice quality. There is also a tendency for women to produce a more standard, or rhetorically correct pronunciation. For example, in Montreal, French Canadian women pronounced the approximant [l] in pronouns and articles such as *il*, *elle*, *la* and *les* more often than men did. Clark and Clark (1990) found that a creaky voice could also reflect

dialect and sex. In a Northern dialect of British English, a creaky voice was observed in over 65% of the syllables for some male speakers.

Studies conducted in the United States and Germany (Herbst, 1969; Takefuta et al. , 1971) showed that females had a greater pitch variability than males. McConnell-Ginet (1974) reported greater pitch range and intonation variability for white, middle-class women in the United States. Other studies (Pellowe and Jones, 1978; Elyan, 1978) also reported that men used a much greater proportion of falling than rising intonation, while women generally used more rising intonation. Furthermore, women displayed a greater variety of intonation patterns than men did. In our experiment on F_0 variation, the standard deviations of F_0s were calculated for 17 male speakers and 17 female speakers over 5 repetitions of the utterance "My name is (speaker's name)", and the results are plotted in Figure 2-4. The average F_0 standard deviation of 17 males is 10.13 Hz; while the average F_0 standard deviation of 17 females is 21.55 Hz. That is to say, the females' average standard deviation is more than double that of the males.

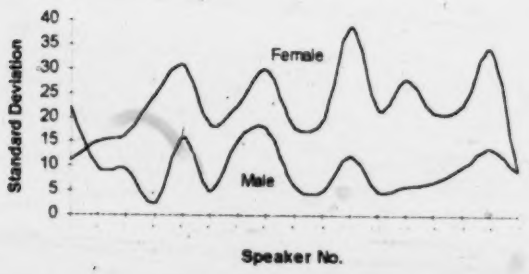


Figure 2-4 Variability of female and male speakers' F_0.

One explanation for the significant F_0 variation difference between male and female is the non-linear characteristics of auditory perception. To produce a perceptually equivalent intonation pattern, a female's F_0 variation has to be larger because a female's natural F_0 is higher than a male's. Furthermore, female

speech is observed to have more intonation variability than a male's speech, which also contributes to greater F_0 variation.

Another interesting phenomenon in the observed sex differences in the phonetic realisation of vowel categories is that the vowels spoken by females exhibit greater between-category dispersion in the F1 x F2 plane than the vowels spoken by males. Different explanations for this phenomenon have been suggested. One interpretation (Sachs et al. , 1973; Ohala, 1984) is that the differences reflect a strategy of the speakers to supplement or exaggerate the acoustic effects of anatomical differences so as to achieve a more masculine-sounding vocal quality in males and a more feminine-sounding vocal quality in females. Another theory (Diehl et al. , 1996) suggests that without the compensatory effect of greater dispersion of vowel categories in the acoustic domain, the typically higher F_0 of female speakers would yield reduced identifiability of vowels because of sparser harmonic sampling of spectral envelopes.

2.4 Speaker Vs. Phonetic Information

There exists distinction between phonetic and speaker information in the acoustic domain. The relative formant positions are an important phonetic cue. It is well documented in the literature that within certain limits, vowels retain their phonemic identity if certain formant frequency ratios are preserved. The absolute formant positions, however, are an important speaker cue. When different speakers produce a set of speech sounds with the same phonetic quality, the relative formant positions will be about the same, but the absolute formant frequency values differ from speaker to speaker.

The low formants (F_1 and F_2) are important for both phonetic and speaker information because they are sensitive to the articulator positioning or movement, as well as to the overall characteristics of a speaker's vocal tract. The high formants, however, are more relevant to speaker information, rather than to the phonetic aspects of the sound because they are mainly decided by the shape of the vocal tract and less disturbed by the articulator positioning or movement.

Fant's study of American English and Swedish Vowels (1973) showed that different vowel classes in both languages had no significant impact on formant F_3. Ladefoged (1982), on the other hand, suggested that "the position of the fourth and higher formants in most vowels is indicative of a speaker's voice quality rather than the linguistics aspects of the sounds" (p.193). Endres et al.'s study (1971) also suggested that the formants in the higher frequency range are more reliable source for a speaker's vocal-tract information, compared with the formants in the low frequency range. In their study, the spectrograms of two professional imitators' imitations were analyzed. They found that the formants in the higher frequency ranges were particularly difficult for mimicry.

Furui (1986) analyzed the long-time speech spectra derived from the averaged cepstral coefficients for a Japanese word / baNgo: / uttered by nine male Japanese speakers. He found that a significant part of the spectral variations among the speakers was above the frequency range of 2.5 kHz; while in the lower frequency range, the spectral envelopes were relatively consistent. This supports the argument that the higher frequency range in the speech spectra contains significant speaker information.

The cross-language long-time speech spectrum comparison performed by Byrne at al. (1994) provides another interesting view of the general speaker-information-distribution pattern in the frequency domain. In Byrne et al.'s experiment, the long-time average speech spectra from speakers of 17 different languages and dialects were compared, using ten male and female speakers for each language or dialect. The speech material was a passage selected from a story book on the basis that it was relatively easy to read and did not involve excessive repetition. The experimental results showed a similarity of the long-time average speech spectra across different languages, and also the inter-speaker variation pattern in the frequency domain. A significant increase in frequency variability existed around 100 Hz for the male groups and around 300 Hz for the female groups. Frequencies around 400 Hz to 1 kHz showed relatively low speaker variability for

both sex groups. An increase of speaker variability began for both groups at around 3 kHz and reached its peak at around 8 to 10 kHz.

Formant transitions or dynamics in the speech signal contain significant phonetic information. For example, the perception of a particular liquid or glide, e.g. [l, r, w, j] depends on the onset frequencies and direction of the formant transitions (Lieberman and Blumstein, 1988). In fricatives, the onset of the noise is fairly gradual. If it is too abrupt, the sound will be perceived as an affricate or a stop (Gerstman, 1957). Formant transitions also provide important cues for the place of articulation in perception of stop and nasal consonants (Lieberman and Blumstein, 1988; Ladefoged, 1975). In speech recognition, using feature parameters associated with formant transitions or dynamics reported significant enhancement of speech recognition performance (Furui, 1986; Chengalvarayan and Deng, 1997).

As for the significance of formant transitions or dynamics for speaker information, Heuvel et al. (1992) found that in Dutch vowels, the transition part contained less speaker information compared with the steady-state one. Heuvel et al. (1995) further investigated speaker variability in the coarticulation of / a, i, u / in /C_1VC_2/ pseudo-words containing the consonants / p, t, k, d, s, m, n, r /. They found that the largest amount of coarticulation was in / u / where nasals and alveolars in C_1 position had the largest effect on the formant positions, especially on F_2. They also found that coarticulation in / a, u / more tended to be speaker-specific. Some researchers (Soong & Rosenberg, 1988; Furui, 1990) used the transitional spectral information in their speaker recognition systems and achieved improved performance.

F_0 is a significant acoustic cue for speaker information. Furui (1986) found in his perceptual experiment that though the F_0, the source spectrum, the formants and the spectral envelope were all relevant to speaker characteristics, the F_0 played the major role in speaker identification. In early speaker recognition research, automatic speaker recognition systems based only on speakers' F_0 information (Atal, 1972) reported a high speaker identification rate. However, the problem

with using only F_0 for speaker recognition is that, F_0 is vulnerable to mimicry. Endres et al. (1971) found that F_0 could be imitated quite accurately by professional imitators. In contrast, the formants, particularly those in the higher frequency ranges, did not correspond with the voice being imitated. Therefore, F_0 may be not as reliable as the formants for computer speaker recognition purpose.

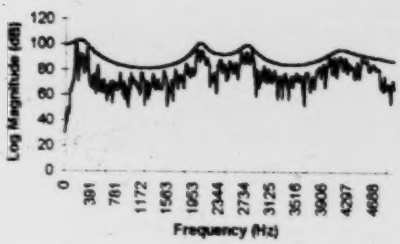
The significance of F_0 for phonetic information, however, is largely language-dependent. In languages like English, F_0 provides phonetic cues for voicing and some suprasegmental features such as stress, intonation and marking of the boundaries of syntactic units. In tone languages, F_0 plays a more significant phonetic role. For register tone languages, the F_0 level carries phonetic information. For contour tone languages, F_0 contours convey different linguistic meaning. There are also tone languages where both the level and contour of F_0 play an important phonetic role (Hogan, 1996).

There were researches attempted to separate speaker information from phonetic information in the speech signal. One method was to use the long-time averaged speech spectrum for speaker information (Furui et al. , 1972). With the averaging of the spectra extracted from each frame of the entire utterance, the phonetic effects were removed from the speech signal. In another study (Furui, 1986), a linear model consisting of phonetic factor, speaker factor and the interaction between phonetic and speaker factors was built and the effects of phonetic, speaker and interaction factors were measured based on the multivariate analysis of variance using χ^2 distributions. The finding was that the phonetic effect was much larger than the speaker effect, and the interaction effect was also relatively large. Furui suggested that the extraction of speaker information was more difficult than that of phonetic information.

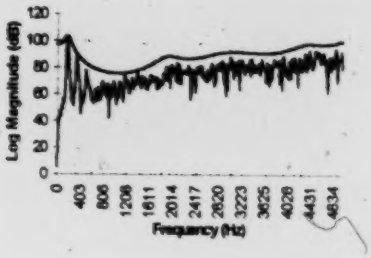
2.5 Intra-Speaker Variation

Intra-speaker variation refers to the normal variability of an individual speaker's voices. As pointed out by Nolan (1983), because of its plasticity, the vocal

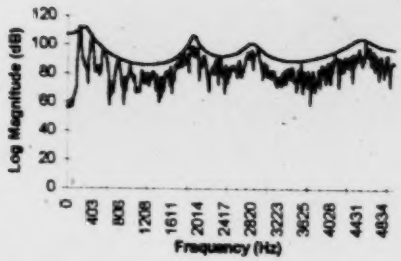
apparatus does not determine particular acoustic characteristics of a person's speech, but merely the range within which variation in a particular parameter is constrained to take place. No one actually repeats the same word exactly the same way even in the same psycho-physical condition. There always exist articulatory differences from trial to trial. Figure 2-5 illustrates the long-time spectra from both DFT and LPC analyses of the six utterances of the same word "zero" spoken by the female speaker F1 from the TI-46 database. These six repetitions were taken from six different recording sessions.



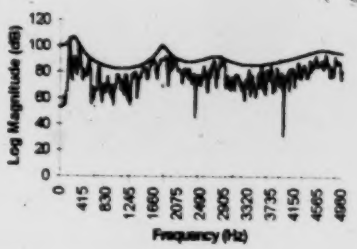
[1]



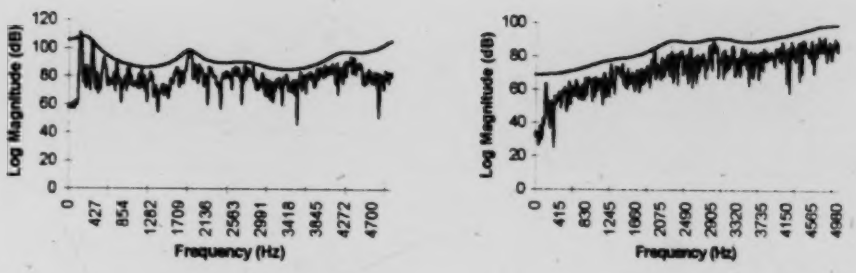
[2]



[3]



[4]



[5] [6]

Figure 2-5 Long-time spectra of six repetitions of the same word "zero" spoken by a female speaker in the TI-46 data. In each graph, the highly varying spectrum is from DFT analysis, and the smooth spectrum from LPC analysis.

Figure 2-5 clearly indicates the spectral differences among six repetitions of the same word spoken by the same speaker. More drastic intra-speaker variation can be caused by fatigue, health condition or emotional status. Another source which contributes to intra-speaker variation is the communicative intent. Nolan (1983) suggests that:

> The speech of an individual will be influenced by the attitude he wishes to convey (affective), by the image he is trying to present (self-presentation) and by the organization he tries to impose on an interaction (interaction management). (p.203)

In speech communication, a speaker uses different intonation, stress, tempo and voice patterns to convey his/her emotion or attitude towards the addressee; s/he may also adjust the speech style or dialect to the social environment. These are what Nolan termed as *affective* aspects of speakers' communicative intent. In conversation, a speaker also tries to present him/herself in a certain way, in terms of masculinity-femininity, self-confidence etc. This is the *self-presentation* aspect of the communicative intent. A very interesting example of both the affective and self-presentation aspects of communicative intent was provided by the speeches and debates by both candidates for the US presidency during last

year's election period on the TV shows. Both candidates used voice strategies in appealing to audiences. *Interaction management* refers to the speech patterns which are used to structure the verbal interaction. For example, a speaker might use an overall intonation pattern to signal the end of his/her speech and yield the "speaking turn" to the other participant.

Intra-speaker variation may also be age-dependent. Research (Sharkey et al. , 1985) found that some of the children's articulatory movements tend to be more variable than adults. This variability was attributed to children's lesser degree of precision in their speech; less habituation of movement patterns; and exploration for motor learning. Smith (1994) showed that young children had longer duration and greater duration variability than adults when producing the same target speech. It was suggested that the duration variability was a useful measure of children's progress toward adult-like speech. However, Stathopoulos (1995) did a study of the acoustic, aerodynamic and respiratory kinematic comparison of children and adults during speech with subjects ranging from 4 to 30 years age, and claimed that the experimental results did not support the argument that children show consistently more intra-speaker variability than adults in their speech.

Intra-speaker variation can also come from the phonetic environment. Whalen and Levitt (1995) found in their study with 31 languages representing 11 of the world's 29 major language families that there is a universal tendency for high vowels to have higher F_0 than low vowels. Moreover, Bruyninckx et al. (1994) found that intra-speaker variation is language-dependent. In their study, bilingual (Catalan/Spanish) speakers were selected in order to neutralize the speaker's individual characteristics. The long-time average spectrum was used as an acoustical measure of intra-speaker variability. They did two kinds of within-speaker variability comparisons: the between-language (Catalan/Spanish) and the within-language (Catalan/Catalan and Spanish/Spanish) ones. It was found that the between-language intra-speaker variability was higher than the within-language one, irrespective of sex and language dominance categories. The data

also show a tendency of greater within-language variability in the dominant language than in the non-dominant one.

With longer time intervals, the intra-speaker variation usually increases. Furui (1986) conducted an investigation on intra-speaker variability with nine male speakers over 15 months. He found in his experimental results that intra-speaker variation increases as a function of the time interval in the first three months. After that, the change is not that significant. For speaker recognition, intra-speaker variation is a serious problem. The speaker model of the speaker recognition system is built on a speaker's training utterances, which are usually collected in a short period of time. If the speaker model has no adaptation ability, then, the performance of the speaker recognition system will deteriorate with time.

In summary, this chapter has discussed speaker information and its representation in the acoustic domain. A speaker's idiosyncratic voice quality is the result of the complicated interaction of his/her anatomical, psychological and social factors. The inter-speaker variations existing in speech signals provide the basis for computer speaker recognition. However, there also exist significant intra-speaker variations, which make the real-world speaker recognition application a very challenging task.

CHAPTER 3
SPEAKER INFORMATION IN THE PHONETIC DOMAIN

Speaker information is embedded in the phonetic environment. In articulation, the involvement of a speaker's vocal organs and articulators varies according to each speech sound. Consequently, different aspects and degrees of speaker information are encoded in a particular speech signal. This chapter examines how speaker information is associated with different phonetic categories.

3.1 Vowels

The production of vowel sounds involves a relatively unobstructed air stream, a relatively large amount of acoustic energy, and a relatively steady-state articulatory position. These features ensure the formation of a relatively stable formant structure. Compared with other phonetic categories, it is expected that vowels contain more speaker information because a stable formant structure provides reliable information on a speaker's glottal-source and vocal-tract characteristics. As for diphthongs, they involve dynamic movements from one articulatory position to another. An interesting question is: does the dynamic articulatory movement acoustically encode more or less speaker idiosyncrasy? O'Shaughnessy (1996) suggested that both the static and dynamic values of the center frequencies of the strongest resonances in speech sound are of great importance to speech perception. Some researchers (Soong & Rosenberg, 1988; Furui, 1990) used the transitional spectral information in their speaker recognition systems and reported improved performance; while Heuvel et al. (1992) found that, in Dutch vowels, the transition part contained less speaker information as compared with the steady-state one.

Within the vowel category, the amount of speaker information still depends on the particular class of vowels. Fant (1973) compared the first three formants of different vowels in both the American English and Swedish data (See Fugue 3-1).

The content in this page has been removed due to copy right restrictions.
The information removed was:

Figure 3-1. The average female/male formant ratios of F1, F2, and F3 from American English and Swedish data. (after Fant, 1973).

The above figure can be found in
Page 85, Fant, G. (1973). Speech sounds and features. Cambridge: The MIT Press.

According to the average male-female formant ratios of F_1, F_2 and F_3 for different vowels of the American English and Swedish data, Fant suggested that there exist quite similar patterns for formant differences between male and female speakers across different vowel categories.

It can be observed from Figure 3-1 that F_1 and F_2 of the rounded back vowels reflect relatively low sex, or vocal-tract-related speaker differences. F_1 of the close or highly rounded vowels is also the same. However, for the very open front or back vowels, the sex difference of F_1 is significantly higher than the average, which Fant called the *sex factor*. Fant's interpretation of the dependency of speaker information on the particular vowel category is that: the low formants (F_1 and F_2) produced with a typical double Helmholtz resonator configuration, such as with rounded back vowels, are less critically dependent on the overall vocal tract length than other formants. The shorter overall vocal tract length of the female speaker can be compensated by narrowing lip-opening and tongue-hump passage. This may explain the relatively low speaker information (or sex difference) in the highly round or close vowels. As for the very open front and back vowels, the vocal cavities behave more like standing wave resonators. Consequently, the characteristics of the vocal tract has a direct effect on all the formants. It may explain the significant increase of speaker information in F_1 in this category of vowels.

Fant's study found no significant difference for formant F_3 across different vowel classes in both languages. This factor may suggest that F_3 is a more reliable source for speaker information, compared with F_1 and F_2.

3.2 Fricatives

Fricatives are characterized by a turbulent air stream which occurs when the air is channeled through a narrow constriction. The spectrum of a fricative sound can be considered as the product of a noise source located supraglottally (except for [h]), which is modified by the front resonator. In the case of voiced fricatives, a second source due to the vibrating vocal folds is also present in the spectrum.

The front resonator involved in the fricative production is usually a small portion of the vocal tract. For example, for fricatives like [s] and [z], the place of articulation involved is the region around the alveolar ridge. Therefore, the derived acoustic parameters may only reflect very limited and partial information of the speaker's vocal tract. Although the part of the vocal tract behind the noise source may contribute anti-resonances to the fricative spectrum, most of the speaker differences in producing fricatives are more likely due to an individual speaker's vocal setting rather than to the glottal source and vocal-tract anatomy.

3.3 Nasal Consonants

Nasals are dominated primarily by the resonances of the nasal-pharyngeal tube and the anti-resonances of the mouth-cavity. The anti-resonances are introduced into the spectrum because the oral cavity acts like a side-chamber which absorbs energy from the nasal cavity. The nasal spectrum varies from speaker to speaker due to variation in the properties of the nasal cavities. Since a speaker's nasal cavity is largely fixed and not likely to be maneuvered during articulation, nasal consonants can be a very good source for speaker information. Su et al. (1974) also suggested that the coarticulation between a nasal and its ensuing vowel contains even more speaker information than the nasal itself. However, nasals are a paradox for computer speaker recognition. Since nasals contain stable speaker information in normal conditions, it is desirable to have more nasal components in a speaker's training and testing utterances. However, the resonances of the nasal cavity are susceptible to changes due to common diseases such as cold and influenza. Congestion of the nasal passages will seriously affect speaker-recognition performance, which causes one serious problem for the real-world application of computer speaker recognition.

3.4 Stops

Stops can be divided into five acoustic segments: *occlusion*, *transient*, *frication*, *aspiration* and *transition* (Fant, 1968). The *occlusion* is the period when the vocal

tract is completely closed and characterised acoustically by the absence of high frequency energy. In the case of voiced stops like [b], [d] and [g], there is a low frequency energy in the 0-500 Hz range, which is sometimes called the "voice bar". The *transient* corresponds to the release of the closure following a sharp rise in intra-oral air-pressure and is acoustically represented by an intense spike of about 10 ms, with energy at all frequencies. The high intra-oral pressure and a narrow opening at the point of release results in *frication*. With an increase in the vocal tract opening, *aspiration*, which is caused by turbulence at the glottis, may arise. The *transition* is the interval from the time at which the formants are first detectable in the aspiration stage to the following vowel formant target. According to Ladefoged's study (1982), there exists speaker variability in the length and type of aspiration that occurs after initial voiceless stops. The rate of transition of the formants after voiced stops is also different from one individual to another. However, the duration of the *transition* segment in a stop sound is usually very short. Listeners may not be able to perceive the individual difference. According to Ainsworth (1976), the onset of a tone does not create an instantaneous sensation of pitch. It usually takes a critical duration before a stable pitch is heard. For tones below 1000 Hz, the critical duration is about 6 $\pm$ 3 cycles, whereas above 1000 Hz, the critical duration is about 10 ms. This minimum threshold duration for tone perception has also been verified in a Mandarin Chinese tone perception experiment (Chen & Rozsypal, 1992). The perceptual limitation on transient sound for human listeners, however, does not necessary hold for the computer. With proper processing of the speech signal, the transient acoustic change may still be detected. However, in the conventional speech and speaker recognition systems, the frame size for short-time spectral analysis (either FFT or LPC) is usually longer than 20 ms, which is inappropriate for catching the very short transition of a stop in the speech signal.

3.5 Quantitative Studies

Quantitative studies of speaker information in the phonetic domain have been extensively reported. Generally speaking, there is consensus on which phonetic categories tend to carry more speaker information. Wolf (1971) found that vowel and nasal spectra were efficient acoustic parameters for speaker recognition. Fakotakis et al. (1993) reported that good speaker verification performance was achieved by using only the vowel segments in the speech signal (91.39% for verification, 90.19% for closed-set identification, 95.28% for open-set identification). Heuvel et al. (1992) found that the rank order of increasing speaker information was: stops + [r], fricatives, short vowels, nasals and long vowels. Floch et al. (1994) reported similar results. They found that vowels, diphthongs and nasals were most favorable for speaker discrimination. Eatock and Mason (1994) produced a phoneme ranking based on the speaker-verification performances. They found nasals and vowels were most informative, followed by fricatives, affricates and approximants. Stops performed the worst.

In summary, this chapter has discussed speaker information in its phonetic environment. Vowels tend to be rich in speaker information because their relatively-stable formant structures provide more reliable information concerning the anatomy of a speaker's glottal source and vocal tract. Within the vowel category, the very open front or back vowels tend to contain more speaker information; while the close or rounded vowels contain less speaker information. Nasals can also be good candidates for spotting speaker idiosyncrasy because of the distinct properties of a speakers' nasal cavity. However, the disadvantage of using nasals in speaker recognition is that the credibility of nasals depends much on a speaker's health condition. The nasal resonances can be easily affected by common illness such as influenza. Other phonetic categories are also discussed concerning their respective importance for speaker information. Quantitative research has supported the close relationship between speaker information and its phonetic environment. In general, vowels, nasals have been proved to contain more speaker information than other phonetic categories.

CHAPTER 4
SPEAKER-INFORMATION EXTRACTION

A speech signal contains both phonetic- and speaker-information elements. During speech recording and signal processing, the environmental and channel noises are also introduced into the speech signal. Consequently, the speech signal under analysis consists of phonetic- and speaker-information elements, as well as noises. Speaker recognition is only interested in the speaker-information elements. The non-speaker-information-related aspects of the speech signal can be potentially confounding factors for identifying or verifying a speaker. How to maximally extract speaker information from the speech signal, then, becomes one of the most challenging tasks in speaker recognition research. In the literature, various methods have been reported on extracting or enhancing speaker information in the speech signal for improving speaker recognition performance.

4.1 Noise-Reduction Approach

This approach enhances speech in the parametric domain by suppressing or eliminating the non-speech elements which are mostly environmental or channel noises introduced during speech recording and signal processing.

Naik and Doddington (1987) introduced the principal spectral components (PSC) as a spectral representation of the speech signal for speaker verification purposes. The basic principle of the PSC approach is as follows: each frame of a speech signal is first processed into LPC coefficients and transformed into a spectral amplitude vector using mel-frequency filter banks. The amplitude vector is then rotated and scaled by the eigen vectors of a covariance matrix, which is estimated by pooling together the spectral amplitude vectors over the entire training data base. The resulting vector consists of uncorrelated features, ranked according to their statistical variance. The least significant features are removed. Those removed features represent mostly invariant environmental or channel

noises. The PSC method reportedly achieved higher performance than other parametric representations for speaker verification (Homayounpour and Chollet, 1994).

Assaleh and Mammone (1994) proposed the adaptive component weighting (ACW) in the cepstral domain. The purpose of ACW is to emphasize the formant structure by attenuating the broad-bandwidth spectral components. According to Assaleh and Mammone, the broad-bandwidth spectral components were found to introduce undesired variability in the linear predictive coding of speech signal, which undermined the speech information extraction.

Rosenberg et al. (1994) applied three different cepstral channel-normalization methods to remove the noises mostly caused by mismatched recording and channel conditions. Their normalization methods treated the non-speech noises as (1) a long time cepstral average; (2) a short time cepstral average; and (3) a maximum likelihood estimate of the cepstral bias. Their experiment showed significant improvement of speaker verification performance with using of the cepstral channel-normalization techniques.

All the above methods enhance speech in the parametric domain by reducing noises or non-speech elements. Since speech consists of both phonetic- and speaker-information. Speaker information is boosted with the enhancement of speech, which explains the improved speaker recognition performances. However, the noise-reduction approach also enhances the phonetic-information elements in the speech signal. As we will illustrate in Chapter 6, the phonetic-information elements in the speech signal can be confounding factors in speaker recognition. This suggests that enhancing speech as a whole can help the extraction of speaker-information, however, it is certainly not an optimal approach for speaker-information enhancement.

4.2 Auditory Approach

This approach is based on human listeners' auditory properties. The fact that human listeners are very effective in "picking up" phonetic or speaker information in a speech signal, fascinated many speech researchers. There were various attempts to simulate a human listener's auditory properties in speech signal processing. One popularly used auditory parameter for speaker recognition is the mel frequency cepstrum coefficients (MFCC). An important aspect of the MFCC is that the bandpass filters designed for cepstral transformation are not linearly spaced in the frequency domain, but based on the "mel" scale, which is a non-linear frequency representation. According to psychophysical studies, the human ear's frequency response is nonlinear above 1 kHz. This led to the mel scale measurement of pitch. The mel scale is defined as the pitch of a 1 kHz tone, 40 dB above the perceptual hearing threshold, for 1000 mels. Other pitch values are obtained by adjusting the frequency of a tone such that it is half or twice the perceived pitch of a reference tone. Davis and Mermelstein (1980) found that the mel frequency representation of the speech signal had significant advantages. Specifically, it gave better suppression of insignificant spectral variation in the higher frequency bands. More detailed description of the MFCC is presented in Chapter 6.

Another example of the auditory approach is the perceptually-based linear prediction (PLP) (Hermansky et al. , 1985; Xu & Mason, 1991), which transforms speech signals from acoustic spectra into perceptual representations according to the human ear's nonlinear transformation of frequency and amplitude of the acoustic signal. Three auditory principles are used in the PLP for estimation of the auditory spectrum: (1) the critical-band spectral resolution; (2) the equal-loudness curve, and (3) the intensity-loudness power law.

As the noise-reduction approach, the auditory approach enhances speech as a whole, instead of speaker information in particular. Different from the noise-

reduction approach, however, the auditory approach enhances speech according to the human's auditory properties.

4.3 Phonetic Approach

The phonetic approach is based on the knowledge that the amount of speaker information coded in the speech signal depends largely on the phonetic environment (see discussion in Chapter 3). This approach enhances speech elements containing those phonetic categories which are believed to be speaker-information rich (such as vowels and nasals). Meanwhile, speech elements containing other phonetic categories will be suppressed, or even eliminated in the speaker-recognition process.

Savic and Gupta (1990) classified speech segments into five broad phonetic categories: Nasal, Voiced 1, Voiced 2, Plosive and Fricative according to their typical spectral differences. A five-state ergodic HMM model was built for each speaker using his/her training speech data. Each state represented a broad phonetic category. To classify the speech frames of an utterance into the corresponding phonetic categories, the Viterbi algorithm was used to obtain the maximum likelihood state sequence which assigned each frame to one of the states (or the phonetic categories). With the classification of all the frames in the training data into five broad phonetic categories, the reference templates and verification thresholds were computed. Each reference template consisted of a mean vector and a covariance matrix. The verification threshold consisted of the Mahalanobis distance and the number of verification votes required to make an 'accept' decision. In the text-independent speaker verification phase, the frames of the testing utterance were classified into the five broad phonetic categories by using the claimed speaker's trained HMM model and matched with the corresponding reference templates. The final verification score was a weighted linear combination of the scores for each individual phonetic category. The weighting for each phonetic category was based on its effectiveness in

discriminating speakers. The experimental results indicated that speaker verification performance was improved by treating the broad phonetic categories separately.

The speaker recognition system based on vowel spotting (Fakotakis et al. , 1993) is another example of the phonetic approach. On the assumption that much of the speaker information was contained in the vowel category, this system extracted only the steady-state vowel parts in the speech signal for the text-independent speaker recognition. Online automatic vowel spotting was based on the energy concentration and the spectral characteristics of the speech signal. This method also reported good speaker recognition results. With training utterances per speaker about 50 sec and the average test utterance duration of 1.3 sec, the average speaker recognition rate was above 90%.

The main problem with the phonetic approach is that the acoustic cues which differentiate speakers with a similar voice are not necessarily be encoded in vowels or nasals, which are usually considered the important phonetic categories for speaker information, but in other phonetic categories. In that case, the speaker recognition system may lose those cues for separating close speakers. The point we try to stress here is that though some phonetic categories tend to contain more speaker information than others do, as far as an individual speaker is concerned, any phonetic category is potentially important for providing his/her individual voice characteristics. Therefore, in speaker recognition, weighting based on phonetic categories is not an optimal approach either. At the same time, there are technical problems for detecting phonetic categories accurately online.

4.4 Speaker-Information Enhancement Approach

In this research, we suggest a new approach for extracting speaker information. Different from the noise-reduction and auditory approaches, this approach will enhance only those speaker-information-related elements in the speech signal, instead of speech as a whole. This approach is also different from the phonetic

approach in that none of the phonetic categories will be treated differently. The only criterion for the weighting strategy of the new approach depends on the distribution pattern of speaker information in the speech parameters, which is estimated from the training speech data. The basic assumption underlining this approach is that a speech signal contains both phonetic- and speaker-information components. Though interrelated, these two kinds of information components have their distinctive representations in the acoustic and parametric domains. For speaker recognition purposes, the optimal way to enhance speaker information is to measure the distribution pattern of speaker information in the acoustic or parametric domain in the training phase, and then apply the corresponding weighting function to enhance only the speaker-information-related elements in the testing phase.

Since the cepstrum is a widely used parametric representation of the speech signal in both speech and speaker recognition applications, the present study focuses on the methods for measurement of speaker information in the cepstral domain, and the optimal weighting strategy for speaker information enhancement. More detailed description of this approach will be presented in Chapter 6 and Chapter 7.

In summary, this chapter has reviewed existing methods in speaker recognition research for extracting or enhancing speaker information. The noise-reduction and auditory approaches enhance speech as a whole, rather than speaker information in particular. As for the phonetic approach, it emphasizes the phonetic categories in the speech signal which tend to contain more speaker information. The problem with this approach is that an individual speaker's voice idiosyncrasy may be coded in those phonetic categories which are usually low in speaker information and consequently be neglected. At the same time, online phonetic-category-spotting is technically difficult.

We propose a new approach for speaker-information enhancement, which focuses on speaker-information elements in the speech signal. The basic method is to measure the distribution pattern of speaker information in the speech

parameters using the training speech data, and in the testing phase, apply the corresponding weighting function to enhance only those speaker-information rich elements in the speech parameters.

CHAPTER 5
CEPSTRAL REPRESENTATION OF THE SPEECH SIGNAL

The proposed new approach for speaker-information enhancement will be tested in the cepstral domain. For a better understanding of the cepstrum, this chapter is devoted to a discussion of some technical aspects of the cepstral representation of the speech signal.

5.1 Parametrization

Parametric representation of the speech signal is done in the front end of computer speech or speaker recognition. Its purpose is to compress the speech data in such a way that the irrelevant elements contained in the speech signal are maximally reduced with minimum loss of relevant information. The choice of a specific parametric representation is determined by the type of speech application and considerations of computational efficiency. In speech recognition, a good parameter should maximally reduce data redundancy with minimum phonetic-information loss. In speaker recognition, however, the parameter should maximally reduce irrelevant elements with minimum speaker-information loss.

5.2 Spectral Representation of the Speech Signal

The spectrum is the basic representation of the speech signal in the frequency domain. Speech is a non-stationary signal. However, it is assumed that the speech signal is stationary in a sufficiently short period of time. Then, either a discrete Fourier transform (DFT) or linear predictive coding (LPC) can be applied to get the short-time spectrum or LPC coefficients for representation of that speech interval. The spectrum is appropriate for speech representation because it conforms with human auditory perception of speech sound. According to auditory perception theory, the basilar membrane within the inner ear responds to the sound wave like a bank of filters. When the incoming sound causes displacements of the basilar membrane, the membrane vibrates at frequencies

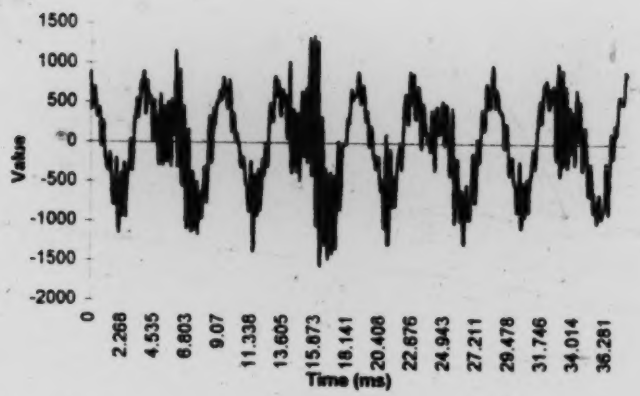
commensurate with the incoming sound frequencies and transmits the frequency information via inner hair cells to higher level auditory nerve systems.

In speech signal processing, the DFT plays a role similar to the basilar membrane. It converts a time-domain signal into a frequency domain representation :

$$\hat{F}(m) = \sum_{n=0}^{N-1} x(n) e^{-j2\pi mn/N} \qquad \text{(m=0, 1, ..., N-1)} \qquad (5.1)$$

where *x(n)* is the sampled sequence; *N* is the number of samples in the short-time analysis interval.

Figure 5-1a illustrates a short interval of a speech signal in the time domain, and Figure 5-1b shows the same speech signal in the frequency domain after DFT. The speech signal was a short interval of the vowel [i] recorded at 16 bits with an 11 kHz sampling rate. The DFT used a Hamming window with duration of 20 ms and a shift of 10 ms.



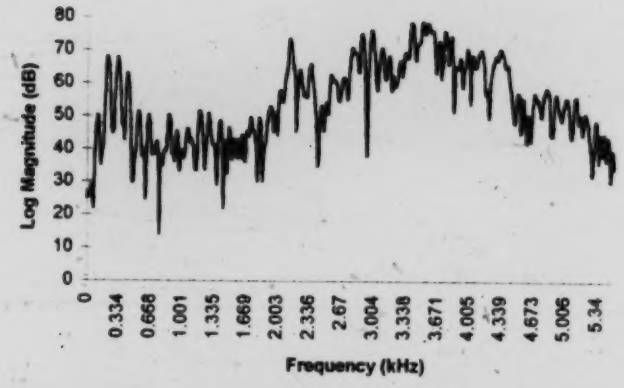


Figure 5-1 (a) Waveform for a short interval of the vowel [i]; (b) DFT of the speech signal.

LPC analysis is also very effective in representing the speech signal. The basic assumption for LPC is that a given speech sample at time *n*, *s(n)*, can be approximated as a linear combination of the past *p* speech samples. The following description of LPC calculation using the autocorrelation method is basically from Rabiner and Juang (1993).

(1) Calculation of autocorrelation coefficients:

$$r(m) = \sum_{n=0}^{N-1-m} x(n)x(n+m), \qquad m = 0, 1, \ldots.p, \tag{5. 2}$$

where *r(m)* is the *m*th autocorrelation coefficient; *N* is the number of samples in the short-time analysis interval; *p* is the order of the LPC analysis.

(2) Converting autocorrelation coefficients to LPC coefficients:

$$E^{(0)} = r(0) \tag{5.3}$$

$$k_i = \left| r(i) - \sum_{j=1}^{i-1} \alpha_j^{(i-1)} r(|i-j|) \right| / E^{(i-1)}, \qquad 1 \le i \le p \tag{5.4}$$

$$\alpha_i^{(i)} = k_i \ , \ \alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)} \ , \quad E^{(i)} = (1 - k_i^2) E^{(i-1)}, \tag{5.5}$$

The set of equations (5.3 - 5.5) are solved recursively for i=1, 2, ..., p, and the LPC coefficients a_m are given as:

$$a_m = \alpha_m^{(p)} \qquad 1 \leq m \leq p. \tag{5.6}$$

LPC provides good spectral approximation of the speech signal, especially for the quasi-stable-state voiced regions of speech. Figure 5-2 is an illustration of spectral representation of a short interval of the vowel [i] (see Figure 5-1a) using LPC analysis.

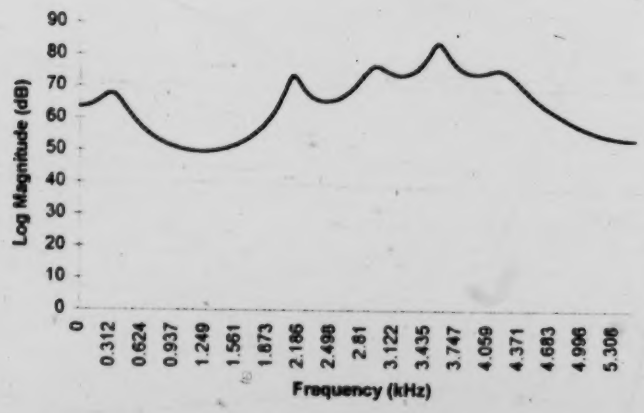


Figure 5-2 Spectral representation of a short interval of the vowel [i] using LPC Analysis.

LPC analysis brings some degree of source-vocal tract separation, which makes the spectrum smoother and the formants more prominent. For better comparison of DFT and LPC spectral representations of the same short speech signal, we re-plotted the DFT spectrum in Figure 5-1b and the LPC spectrum in Figure 5-2 as Figure 5-3.

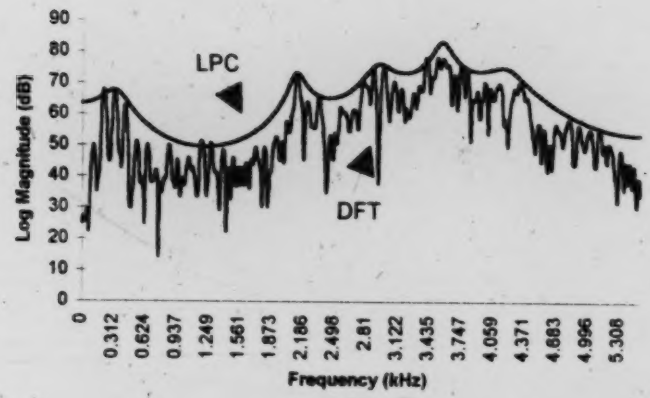


Figure 5-3 DFT and LPC spectral representations of a short interval of the vowel [i].

Both DFT and LPC are used extensively in speech signal processing. However, there are inherent problems with these two spectral representations for speaker recognition purposes. The disadvantage of LPC is that it is an all-pole model, which overlooks the anti-resonances (zeros) in the speech signal. The anti-resonances, as we have discussed in Chapter 2, contain significant speaker information especially for nasal sounds. As for the DFT, it can be observed from Figure 5-3 that there are rapidly varying components which are superimposed on the spectral envelope. These rapidly varying components occur because of the interaction between the vocal tract transfer function h(t) and the quasi-periodic excitation glottal source g(t). The voiced speech signal x(t) can be seen as the convolution of g(t) and h(t):

$$x(t) = \int_0^t g(\tau)h(t-\tau)d\tau \qquad (5.7)$$

or

$$X(\omega)=G(\omega)H(\omega) \qquad (5.8)$$

where $X(\omega)$, $G(\omega)$, and $H(\omega)$ are the Fourier transforms of $x(t)$, $g(t)$ and $h(t)$, respectively. The problem with the convolution of the glottal source and vocal tract transfer functions in the spectrum is the mix-up of glottal and vocal tract information. The strong presence of a rapidly varying component due to the periodic excitation obscures the formant information. This is undesirable for speaker recognition because the formant positions indicate a speaker's vocal-tract characteristics, and the glottal source spectrum also contains significant speaker information about the speaker's subglottal and laryngeal status. To separate the glottal source from the vocal tract transfer function, then, is advantageous for speaker recognition.

5.3 Homomorphic Filtering

Homomorphic filtering is a class of nonlinear signal processing techniques that is based on a generalization of the principle of superposition that defines linear systems (Schafer and Rabiner, 1990). In a conventional linear system, signals are composed of added components. The use of linear operators can easily separate them. In a nonlinear system, however, signals are combined by multiplication and convolution. A conventional linear operator cannot be directly applied to separate the component parts. In that situation, homomorphic filtering becomes necessary to convert the nonlinearly combined signals to the linear domain, so that the signal can be treated with conventional techniques. The cepstrum is a kind of homomorphic process, and it is defined as the inverse Fourier transform of the logarithm of the power spectrum of a signal (Rabiner and Schafer, 1978). The cepstrum is appropriate for separating the source $G(\omega)$ and vocal transfer function $H(\omega)$ because these two components are combined by convolution in the speech signal . The cepstral analysis transforms $G(\omega)$ and $H(\omega)$ into a summation:

$$\log|X(\omega)|=\log|G(\omega)| +\log|H(\omega)|, \qquad (5.9)$$

$$c(i)=F^{-1}\log|X(\omega)|=F^{-1}|\log|G(\omega)|+F^{-1}|\log H(\omega)|. \quad (5.10)$$

The glottal source and the vocal tract transfer function are separated in the quefrency domain. The peak interval in high quefrency represents the glottal source and the vocal tract resonance is represented in low quefrency region.

The cepstral coefficients used for parametric representation of speech can be derived from the discrete cosine transform of the log filter bank outputs:

$$c_i = \sum_{j=1}^{p} m_j \cos\left(\frac{\pi i}{p}(j-0.5)\right), \quad 1 \le i \le N. \quad (5.11)$$

where c_i is the ith cepstral coefficients; p is the analysis order; m_j is the jth log filter bank output resulted from the DFT and filter bank analysis; and N is the defined number of cepstral coefficients.

The cepstral coefficients can also be derived from the LPC coefficients (Rabiner and Juang, 1993):

$$c_0 = \ln\sigma^2, \quad (5.12a)$$

$$c_m = a_m + \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k}, \quad 1 \le m \le P, \quad (5.12b)$$

$$c_m = \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k}, \qquad m>P. \quad (5.12c)$$

where σ^2 is the gain term in the LPC model; a_m is the mth LPC coefficient; P is the number of LPC coefficients.

The lower-order cepstral coefficients are related to the global pattern of the logarithmic spectrum, whereas the higher-order ones are more related to spectral details (Furui, 1986). The cepstrum has the advantage of being invariant to fixed spectral distortions from recording and transmission environments (O'Shaughnessy, 1987). Davis and Mermelstein (1980) conducted an experiment

with different parametric representations of the speech signal. They found that the cepstral parameters such as linear frequency cepstrum coefficients (LFCC), mel-frequency cepstrum coefficients (MFCC) and linear prediction cepstrum coefficients (LPCC) performed better than other parameters in capturing significant speech information. Davis and Mermelstein suggested that a Euclidean distance metric defined on the cepstrum parameters gives better separation of phonetically distinct spectra. The representation of acoustic information in the hyperspace of the cepstrum parameters favours the use of a particularly simple distance metric.

However, some inherent problems with the cepstral analysis need be taken into account. As discussed by Deller et al. (1993), the cepstrum may fail to resolve the low- and high-quefrency components and may have the potential of improperly emphasizing the low-level noise portions of the speech spectrum. For people accustomed to think in terms of frequency, quefrency may also cause confusion.

In summary, this chapter has discussed the cepstral representation of the speech signal. According to the source-filter theory, the speech signal is the convolution of the glottal source spectrum and the vocal tract resonances. In speaker recognition, it is desirable to separate the glottal source from the vocal tract transfer function because they are encoded with different aspects of a speaker's anatomical characteristics. The cepstrum is a kind of homomorphic filtering which separates the glottal source spectrum from the vocal tract resonant frequencies in the quefrency domain. Consequently, it provides a better spectral representation of the speech signal for speaker recognition purposes.

CHAPTER 6

SPEAKER-INFORMATION DISTRIBUTION IN THE CEPSTRAL DOMAIN

In the previous chapters, we have discussed speaker-information representation in the acoustic and phonetic domains, and the distinction between *speaker* and *phonetic* information in the speech signal. On the bases of these discussions, we have proposed in Chapter 4 a new approach for speaker-information enhancement. Different from the existing speaker-information extraction methods which either enhance speech as a whole, or enhance speech elements associated with certain phonetic categories (see discussions in Chapter 4), the new approach suggests to enhance only those speaker-information-related elements in the speech signal.

In order to develop an optimal strategy for enhancing the speaker-information-related elements in the speech signal, we first need to find out the distribution pattern of speaker information in the speech parameters. In this chapter, we investigates speaker-information representation in the cepstral domain.

6.1 Experimental Design

An experiment is designed to perform quantitative measurements of speaker information coded in mel frequency cepstrum coefficients (MFCC) (Davis & Mermelstein, 1980). The basic assumption of this experiment is that the variance in each MFCC coefficient contains different degrees of speaker information, which may contribute to the separation of one speaker's voice from the other's. To estimate the amount of speaker information coded in each MFCC coefficient for a particular speaker, then, we can intentionally exclude the variance in a MFCC coefficient in turn from a pattern distortion measurement and see how it will statistically affect the separation of this speaker from the rest of the speakers

in the database. The degree of its effect is used as an indicator of the amount of speaker information coded in that particular MFCC coefficient.

Two different statistical methods are used in the present experiment for measuring speaker information. One is *the inter-distribution distance measurement* (Homayounpour & Chollet, 1994), which measures the change of statistics in the intra- and inter-speaker distortion score distributions when the variance in a particular MFCC coefficient is excluded from the distortion distance measurement. The other is *the speaker identification error rate measurement*, which measures the change of the speaker identification error rate when the variance in a particular MFCC coefficient is excluded. Detailed descriptions of both methods will be presented in the following relevant sections. The purpose of using two different statistical methods for speaker-information measurement is to compare which method provides a better estimation of the speaker-information distribution pattern in the MFCC coefficients.

To validate the hypothesis that speaker and phonetic information have their distinctive distribution patterns in the parametric domain, the distribution of phonetic information in the MFCC coefficients is also investigated. The phonetic-information distribution in the MFCC coefficients is estimated by using *the speech recognition error measurement*. Similar to *the speaker identification measurement*, this method measures the change of the speech-recognition-error rate when the variance in a particular MFCC coefficient is excluded from the distortion measurement.

6.2 Speech Database

This experiment needs two sets of speakers' data. One set of data is for training speaker models; the other set of data is for applying either *the inter-distribution distance measurement* or *the speaker identification error rate measurement* to estimate the amount of speaker information coded in each MFCC coefficient.

The speech database we used in the experiment is TI-46 Speech Data, which was designed and collected at Texas Instruments (TI) in 1980 (Doddington & Schalk, 1981). This database was originally designed for the purpose of testing speech recognition systems. It is available on CD-ROM by the National Institute of Standards and Technology (NIST) and is distributed with the permission of Texas Instruments. TI-46 consists of two sets of vocabulary: TI-ALPHA and TI-20. The TI-ALPHA vocabulary contains twenty-six English alphabets. In Chapter 2, we used some of its data for illustration of the inter- and intra-speaker spectral variations. In the present investigation of the distribution pattern of speaker information, we only use the TI-20 vocabulary. The TI-20 vocabulary list includes the following twenty words:

TI-20 Vocabulary List

[1]. zero [2]. one [3]. two, [4]. three [5]. four

[6]. five [7]. six [8]. seven [9]. eight [10]. nine

[11]. enter [12]. erase [13]. go [14]. help [15]. no

[16]. rubout [17]. repeat [18]. stop [19]. start [20]. yes

The TI-20 data corpus contains speech from 16 speakers: 8 male speakers labelled M1 to M8 and 8 female speakers labeled F1 to F8. There are nine recording sessions for each speaker. The first session has 200 tokens, 10 repetitions for each word. The words were collected in rotation, that is, the speaker read the word list 10 times, rather than repeating each word 10 times at once. This session was originally designated for speech model training (or enrollment). The other eight sessions were originally designated as speech testing sessions. Each session recorded 40 tokens in a different random order, with 2 repetitions for each word. The data collection stretched out for nearly two

months. Figure 6-1 illustrates the percentages of each broad phonetic class contained in the phonemic inventory of the TI-20 vocabulary:

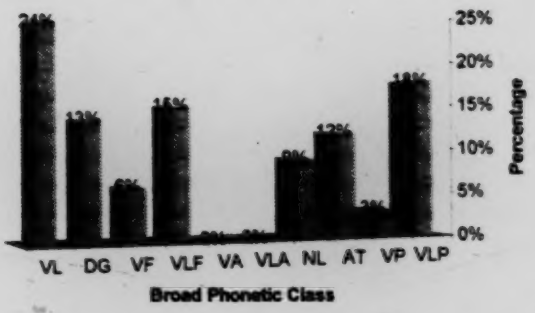


Figure 6-1 Phonetic composition of the TI-20 vocabulary, where VL stands for vowel; DG: diphthong; VF: voiced fricative; VLF: voiceless fricative; VA: voiced affricate; VLA: voiceless affricate; NL: nasal; AT: approximant; VP: voiced stop; VLP: voiceless stop.

These data were recorded in a low-noise sound-isolation booth, using an Electro-Voice RE-16 cardoid dynamic microphone, positioned two inches from the speaker's mouth and out of the breath stream. The data were sampled at 12-bit precision with a 12.5 kHz sampling rate, and stored in NIST-wave format.

The reason for selecting the TI-20 speech database is that this study needs sufficient word repetitions. As we have mentioned, two sets of data are required for this experiment: one set for speaker model training and the other set for speaker-information estimation. We also need a third set of data in the next experiment (Chapter 7) for testing purposes. In the TI-20 database, each speaker has 26 repetitions for every word item from 9 separate recording sessions, which provides an adequate number of sessions and tokens for the present study. The data assignment in this study is as follows: Session 1 is designated to the speaker model training; Sessions 2-5 are used for the measurement of speaker

information; Sessions 6-9 are used for testing different weighting approaches (Chapter 7).

Since the TI-20 data were originally designed for speech recognition purposes, the vocabulary consists mostly of single-syllable words. The average length of each word is about 300 ms in duration. This utterance length is not long enough to yield high speaker recognition performance. However, our study is interested in maximally extracting speaker information from a limited speech source. The data provide the necessary challenging environment.

6.3 Data Pre-Processing

The digitized speech wave files in the TI-20 data contain pre- and post-word silences. For better speaker-information measurement, the silences need to be removed before parametrization. A simple energy-based word-detection algorithm was used to automatically remove the silences. The word-detection program checks the energy level in a 25.6 ms moving window. Two preset energy thresholds are used to detect the word boundaries. The Low Threshold (LT) is determined by the noise level of the recorded signal, and the High Threshold (HT) is empirically set to three times that of the LT (HT = 3 * LT). If the energy level in the current window is higher than the LT, then the frames of the recording in the wave file start being copied into a new buffer. If the energy level drops below the LT in the search process, the buffer will be cleared, and the frame copying will start from the beginning of the buffer again when the energy level rises higher than the LT. This search continues until the energy level in subsequent frames is higher than the HT. Then the word-detection algorithm supposes that the initial part of the speech signal has been spotted with high confidence, and the frames of the recording already being copied to the buffer will no longer be overwritten. The frame copying continues until the algorithm finds that the window energy level is lower than the LT, and a continued search for a certain duration still cannot find any frame whose energy level is greater than HT. Then, the end of

the speech signal is presumably found and the word-detection completed. The frames copied in the new buffer will be saved as the end-pointed speech for parametrization (see the next section). Since the automatic word-detection used here is totally energy dependent, it is possible that the initial and final consonant, expecially stops and fricatives, might be lost because they are usually at very low energy level.

6.4 Parametric Representation of the Speech Signal

The MFCC is used for parametric representation of speech signals. The block diagram of the MFCC processing is illustrated in Figure 6-2

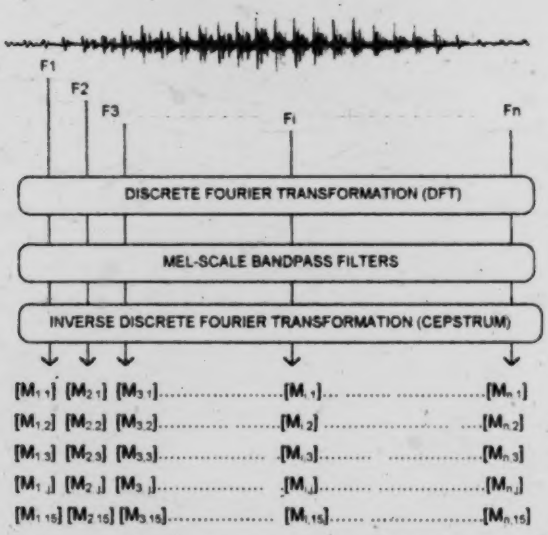

F1
F2
F3
Fi
Fn
DISCRETE FOURIER TRANSFORMATION (DFT)
MEL-SCALE BANDPASS FILTERS
INVERSE DISCRETE FOURIER TRANSFORMATION (CEPSTRUM)
[M1,1] [M2,1] [M3,1] ..... [Mi,1] ..... [Mn,1]
[M1,2] [M2,2] [M3,2] ..... [Mi,2] ..... [Mn,2]
[M1,3] [M2,3] [M3,3] ..... [Mi,3] ..... [Mn,3]
[M1,j] [M2,j] [M3,j] ..... [Mi,j] ..... [Mn,j]
[M1,15] [M2,15] [M3,15] ..... [Mi,15] ..... [Mn,15]

Figure 6-2 Block diagram of the MFCC processing.

In Figure 6-2, F_i represents the *i*th speech frame, which is 25.6 ms long with a frame shift of 10 ms, and is weighted by a Hamming window. M represents a 15 dimensional vector. M_{ij} is the jth MFCC coefficient of the ith speech frame. The

reason for preference of a DFT approach to LPC in this study is that LPC is an all-pole model. The significant speaker information coded in the anti-resonances of the spectrum, such as in cases of nasals, may be lost by the LPC analysis. In addition, LPC was found to provide an inaccurate representation of the consonantal spectra (Davis & Mermelstein, 1980).

In the spectral analysis, mel frequency is used instead of a linear-frequency scale because the non-linear frequency representation, as we have discussed in Chapter 4, is closer to human perception. The mel frequency is calculated directly from the output of the DFT as:

$$\text{Mel}(f) = 2595 \log_{10}(1+f/700). \qquad (6.1)$$

The filters used are triangular (bandwidth = 110 mels, spacing = 55 mel) and they are equally spaced along the mel-scale as shown in Figure 6-3.

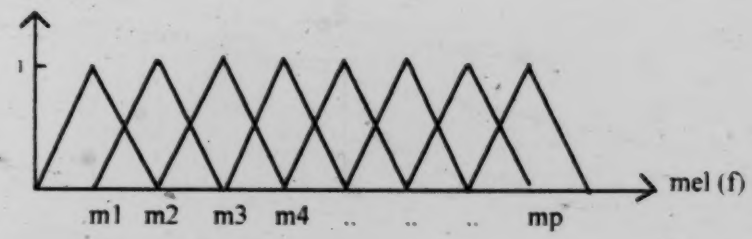


Figure 6-3 Mel-Scale Filter Bank.

where p is the number of triangular bandpass filters. In this experiment, p is set to 20. The MFCC coefficient c_i is calculated by the discrete cosine transform of the log filter bank outputs m_j.

$$c_i = \sum_{j=1}^{p} m_j \cos\left(\frac{\pi i}{p}(j-0.5)\right), \quad 1 \le i \le N \qquad (6.2)$$

where p is the number of mel-scale bank filters, and N is the required number of cepstrum coefficients. In this experiment, N is set to 15.

6.5 Speaker Modelling

For estimation of speaker information in the MFCC coefficients, we first need to train the speaker models for each word item. These speaker models will be used later for distortion measurement. There are two different statistical modeling techniques: *parametric* and *nonparametric*. Parametric modelling, such as Hidden Markov Models (HMM), is based on certain assumptions about the distribution of the parameters of the sampled population, while nonparametric modeling, such as Vector Quantization (VQ), makes no pre-assumptions. In our experiment, VQ is used for speaker modeling, which has the advantage of modeling complex vector spaces with arbitrary precision by simply designing a sufficiently large code book (Picone,1990). A codebook representing a speaker's reference template can be generated by clustering the feature vectors of that speaker's training data. This technique has been proven to be quite efficient for characterizing speaker-specific features (Furui, 1994). There are different algorithms for building a VQ codebook. The method used here is the binary splitting algorithm or LBG algorithm (Linde et al. , 1980; Rabiner and Juang, 1993). The detailed binary splitting algorithm can be described as follows:

(a) Initialization. Set **m** = 1 and calculate the global centroid (codeword) $\hat{c}_1$ for the entire speech training sequence {x_j; j = 1, ...,n}.

(b) Given the $\hat{c}_m$ containing **m** code words {$\hat{c}_i$; i = 1, ...,m}, split each code word $\hat{c}_i$ into two close code words $\hat{c}_i(1+\sigma)$ and $\hat{c}_i(1-\sigma)$, where σ is a

fixed small number. Typically σ is chosen in the range $0.01 < \sigma < 0.05$. $\tilde{c}_{(m)}$ now has 2 * m code words. Assign D' = 0;

(c) Assign each speech frame, which is represented by a vector of 15 MFCC coefficients, to the closest codeword of the current codebook, then update the code word by calculating the centroid of all the training vectors which have been assigned to it;

(d) Compute the average distortion of the training data with the updated codebook:

$$D = \frac{1}{n}\sum_{j=1}^{n} \min_{1 \le i \le m} d(x_j, \tilde{c}_i) \qquad (6.3)$$

where $d(x_j, \tilde{c}_i)$ is the distortion between a training vector x_j and a centroid vector of the updated jth codeword. The distortion is calculated by the Euclidean distance measure.

If $(D - D') > \delta$ (the preset threshold), set D' = D and iterate Step (c) to continue the nearest-neighbor search; if $(D - D') \le \delta$ and m is smaller than the predefined codebook size M, then go to Step (b) to continue the code splitting process;

(e) Iterate Step (b), (c), and (d) until a codebook of size of M reached.

The block diagram of VQ codebook training is illustrated in the following figure.

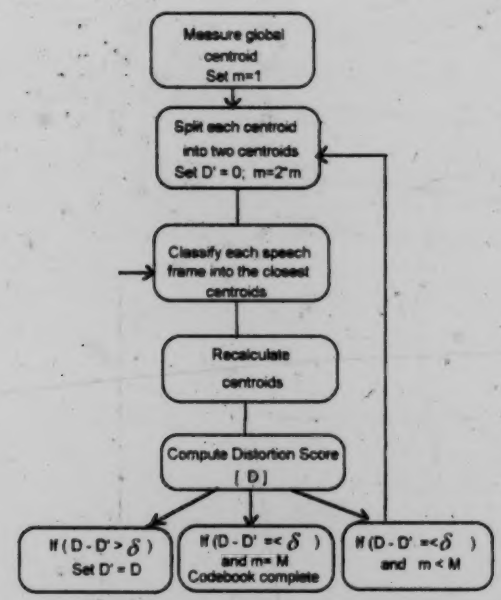

Measure global centroid
Set m=1
Split each centroid into two centroids
Set D' = 0; m=2*m
Classify each speech frame into the closest centroids
Recalculate centroids
Compute Distortion Score (D)
If (D - D' > δ) Set D' = D
If (D - D' =< δ) and m= M Codebook complete
If (D - D' =<δ) and m < M

Figure 6-4 Block diagram of VQ codebook training.

In this experiment, the VQ codebook size is set to 32. The choice of codebook size depends on the phonetic contents of the training data. In general, larger codebook size provides better resolution with the cost of increased computation time. The VQ codebook built for the present experiment is word-based. Since all the words in the TI-20 data are very short (one or two syllables, and no more than five phonemes), a 32-codeword VQ codebook seems adequate for covering the significant spectral categories.

The first recording session of the TI-20 speech data was used to build speakers' VQ codebooks. Since the vocabulary consists of 20 words, twenty codebooks were built for each speaker. Each codebook was trained on the 10 repetitions of a word. The total number of codebooks built for 16 speakers was 16 x 20 = 320.

6.6 VQ Distortion Score Measurement

Since in both *the inter-distribution distance measurement* and the *speaker identification error measurement*, we need to calculate the VQ distortion scores for each testing utterance, the algorithm used for VQ distortion measure is described here. The method for calculating distortion between a speech vector and a codeword in a VQ codebook is the *variance-weighted Euclidean distance measure* (Tohkura, 1986).

$$d(x,x') = \sum_{i=1}^{p} w_i |xi - x'i|^2 ; \qquad (6.4)$$

where d is the distance score; p is the number of cepstral coefficients; w_i is the inverse of the ith cepstral coefficient variance.

The reason for using inverse-variance weighting of the MFCC coefficients in this experiment is that there exist large differences among the variances of different MFCC coefficients. Figure 6-5 shows the variances of the 15 MFCC coefficients obtained from the TI-20 Session 1 speech data. The low-order coefficients obviously have much larger variances than the high-order ones. If the MFCC coefficients were not properly weighted, then the variances contained in the lower coefficients would play a dominant role in the VQ distortion measurement. It has been reported that the weighted cepstral distance measure got significantly better results in both speech recognition and speaker verification experiments (Paliwal, 1982; Nikaus et al. , 1983; Furry, 1981; Tohkura, 1986).

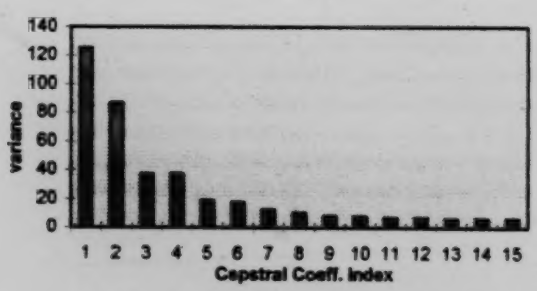


Figure 6-5 Variance distribution in the MFCC coefficients.

In VQ distortion measurement of a testing utterance, each speech frame is 25.6 ms long and represented by a vector of 15 MFCC coefficients. For each speech vector, the closest codeword is searched based on the Euclidean distances between the vector and each of the 32 codewords of the speaker's codebook. Then, the distortion scores of each vector (or speech frame) against its closest codeword are added together across the whole utterance. The final VQ distortion score for a testing utterance is the average score of all the frames of the utterance:

$$D = \frac{1}{n}\sum_{t=1}^{n} d(x_t, \hat{x}_t), \tag{6.5}$$

where

$$\hat{x}_t = \arg\min_{y_i \in C} d(x_t, y_i). \tag{6.6}$$

6.7 Inter-Distribution Distance Measurement

The inter-distribution distance (IDD) measurement is one the two statistical methods which we used in this study to estimate speaker information in the MFCC coefficients. This method was originally proposed by Homayounpour and Chollet (1994) for evaluation of speaker recognition performance. The basic

assumption of IDD measurement is that any improvement in speaker recognition performance will require a wider separation of the intra- and inter-speaker distortion score distributions. The IDD score will increase if the distance between the two means of the inter- and intra-speaker distributions is getting larger and if the inter- and intra-speaker standard deviations are getting smaller. Figure 6-6 illustrates a case of intra-speaker and inter-speaker VQ distortion score distributions:

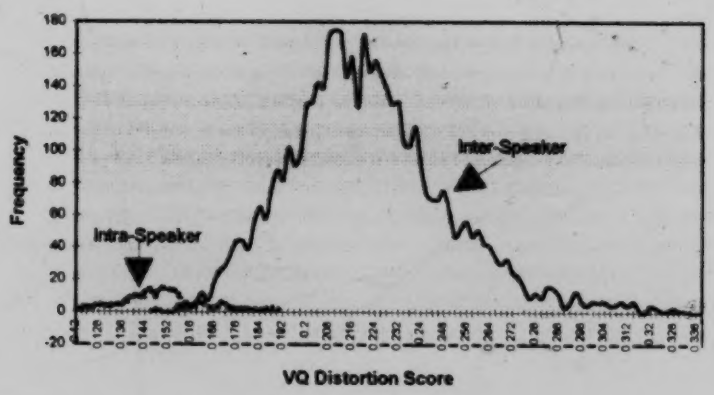


Figure 6-6 A case of intra-speaker and inter-speaker VQ distortion score distribution pattern.

The algorithm used in the present experiment is an adapted version of Homayounpour and Chollet's, to make it statistically more robust for speaker information measurement.

$$IDD(i) = \frac{\mu_2 - \mu_1}{\sqrt{\frac{\sum(d_{i2} - \mu_2)^2}{n_2 - 1}} + \sqrt{\frac{\sum(d_{i1} - \mu_1)^2}{n_1 - 1}}} \qquad (6.7)$$

where *IDD(i)* is the *ith* speaker's IDD score; μ_1 is the mean of the intra-speaker scores; μ_2 is the mean of the inter-speaker scores; d_{i1} is the speaker's *i*th testing utterance's VQ distortion score; d_{i2} is impostors' *i*th testing utterance's VQ distortion score. "Impostors" here are defined as all the speakers in the database other than the speaker whose VQ speaker model is currently being used for distortion measurement; n_1 is the number of the speaker's total testing utterances; n_2 is the number of the impostors' total testing utterances.

To measure an individual speaker's IDD score, two sets of VQ distance scores are necessary. One set is the intra-speaker distortion scores, which are the distance scores obtained when a speaker's utterances are matched with his/her own corresponding word VQ codebooks. The other set is the inter-speaker distortion scores, which are the distance scores obtained when the utterances of the rest of the speakers in the database are matched with that speaker's corresponding word VQ codebooks.

The computer program used in the present experiment has been designed in such a way that it automatically loads each speaker's codebooks in turn and measure the VQ distortion only with its corresponding utterances of all the speakers in the testing data. The computer program also identifies each utterance's VQ score as either an intra- or inter-speaker distance score and save it accordingly. The identification of each utterance's speaker ID and utterance ID by the computer program is based on the wave file name. In the present experiment, Session 2 to Session 5 of the TI-20 data are used for IDD score measurement. There are 20 words with 8 repetitions for each speaker. Consequently, there are 160 intra-speaker VQ distortion scores and 2400 inter-speaker VQ distortion scores for each speaker. Based on these two sets of scores, the IDD score for each speaker can be calculated by using the IDD algorithm.

For estimation of speaker information in each MFCC coefficient, we first conduct a baseline experiment. The baseline calculates the IDD scores for each speaker,

with all the 15 MFCC coefficients included in the intra-and inter-speaker VQ distortion measure. Then, we repeat the same procedure except that, in turn, one of the 15 MFCC coefficients is excluded from the intra- and inter-speaker VQ distortion measure. The purpose is to prevent the variance contained in that particular MFCC coefficient from contributing to the Euclidean distance. The individual speakers' IDD scores may be affected when one of the MFCC coefficients is being excluded from the distortion measurement. Increase or decrease of the IDD scores depends on the amount of speaker idiosyncrasy contained in that particular coefficient. There are three possible cases:

- If the variance in a MFCC coefficient contains significant speaker information, the IDD score will decrease when that coefficient is excluded.
- If the variance in that MFCC coefficient contains little speaker information, the IDD score basically remains unaffected when that coefficient is excluded.
- If the variance in that MFCC coefficient contains significant confounding variation for speaker recognition, the IDD score will increase when that coefficients is excluded.

Consequently, the amount of speaker information contained in each MFCC coefficient can be approximated by the difference between the baseline IDD score, which includes all the coefficients in its VQ distortion measurement, and the IDD score which is measured with one of the MFCC coefficients excluded. The resulting difference score is introduced here as *the normalized inter-distribution distance score* (NIDD). We need to stress here that the NIDD score can be either positive or negative. The higher the NIDD score, the more speaker information is contained in that particular coefficient. A negative NIDD score means that the variance contained in that particular coefficient is mostly confounding for speaker recognition. In the subsequent discussions, we will refer to that kind of variance as *negative speaker information*.

The average NIDD score of all the speakers is defined as:

$$\overline{NIDD}_j = \frac{1}{N} \sum_{i=1}^{N} NIDDij \,. \tag{6.8}$$

where $\overline{NIDD}_j$ is the average NIDD score for the *j*th MFCC coefficient; *N* is the total number of speakers; $NIDD_{ij}$ is the NIDD score for *i*th speaker's *j*th MFCC coefficient.

Experimental Results

There are a total of 16 IDD measurements performed, which include the baseline IDD and the 15 IDDs in which each MFCC coefficient is excluded from the intra- and inter-speaker VQ distortion score measure in turn. The resulting IDD scores are normalized into NIDD scores. The average NIDD score distribution in the MFCC coefficients over all the speakers is presented in Figure 6-7; the average NIDD score distribution over the female group is presented in Figure 6-8; the average NIDD score distribution over the male group is presented in Figure 6-9; the individual speakers' NIDD scores are presented in Appendix 1-A, and their graph representations are presented in Appendices II-A and II-B.

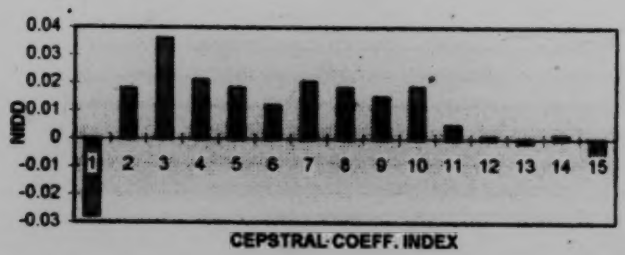


Figure 6-7 Average NIDD score distribution in the MFCC over all the speakers.

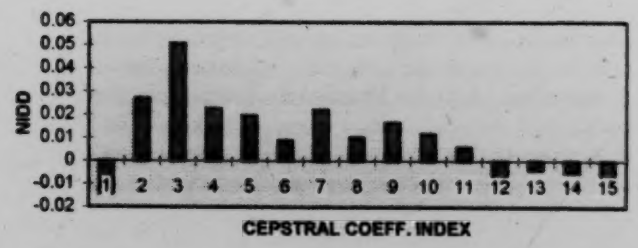


Figure 6-8 Average NIDD score distribution in the MFCC over the female speaker group.

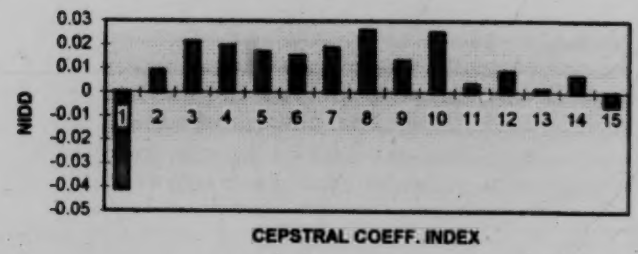


Figure 6-9 Average NIDD score distribution in the MFCC over the male speaker group.

The average NIDD score distribution in the MFCC over all the speakers (Figure 6-7) reveals that the lowest-order coefficient (C_1) and the high-order cepstral coefficients (C_{12} and above) contain little or even negative speaker information. Most of the speaker information is concentrated in the lower- and middle-order coefficients.

As for sex differences, a comparison of Figure 6-8 and Figure 6-9 shows that the high-order coefficients (C_{12} and above) are all characterized by negative NIDD values for the female speaker group; while for the male speaker group, the coefficients in that region mostly have low, but still positive, NIDD values. This suggests that there may exist some kind of different distribution patterns of speaker information in the MFCC coefficients between female and male speakers.

Since Figures 6-7, 6-8 and 6-9 are based on the average NIDD scores of the speaker groups, they only reflect some general tendencies of the distribution pattern of speaker information in MFCC. For a closer examination, we need to look at each individual speaker's NIDD score distribution pattern. Appendices II-A & II-B show the graphs of each individual speaker's NIDD score distribution in the MFCC. It can be observed from these graphs that speaker-information distribution in the cepstral domain, though it may be subject to certain general tendencies, is largely speaker-dependent. Each speaker has his/her unique distribution pattern of speaker information in the MFCC. A certain coefficient can have significantly high NIDD score for one speaker, but less so for another speaker. An individual speaker's NIDD score distribution may also not be consistent with the average distribution pattern of speaker information. Further discussion will be presented in the discussion part of this chapter.

6. 8 Speaker Identification Error Rate Measurement

Another approach for speaker-information estimation is the speaker identification error rate measurement (SIER). The SIER uses speaker identification tests to estimate speaker information contained in each cepstral coefficient.

Speaker recognition can be classified into two categories: *speaker verification* and *speaker identification*. Speaker verification verifies whether the speaker is truly the person claimed to be by comparing his/her utterance with the model of the speaker claimed to be. If the distortion score is within the preset threshold, the speaker will be accepted, otherwise s/he will be rejected.

Speaker identification identifies who is speaking from a known population. There are two types of speaker identification: *closed-set* identification and *open-set* identification. For *closed-set* identification, the utterance is assumed to be spoken by a speaker within the known population. The utterance is matched with all the speaker models and the system identifies the utterance with the speaker whose

speaker model yields the least distorted score (or the highest likelihood score). For open-set identification, the speaker may not be in the known population. Therefore, finding the closest speaker model does not necessarily mean that the speaker's identity is verified. A further verificaton process is used to accept/reject the speaker with a preset threshold. In this respect, the open-set identification is a hybrid method, which includes both speaker verification and identification.

Speaker recognition can also be classified as *text-dependent* or *text-independent* according to whether the speaker is restricted to say a predefined utterance or not. For text-dependent speaker recognition, the speaker model is built on a particular utterance, and in the testing phase, the speaker must say that utterance. In text-independent speaker recognition, the speaker model is built on a training data which include the whole phonetic inventory of the target language. In the testing phase, the speaker is free to say any utterance.

The present SIER measurement uses the closed-set text-dependent speaker identification test. The same VQ codebooks built from Session 1 of the TI-20 data for the IDD measurement are still used here as text-dependent speaker models. Sessions 2 to 5 of the TI-20 data are used as testing data. The speaker identification error measurement program is illustrated in Figure 6-10.

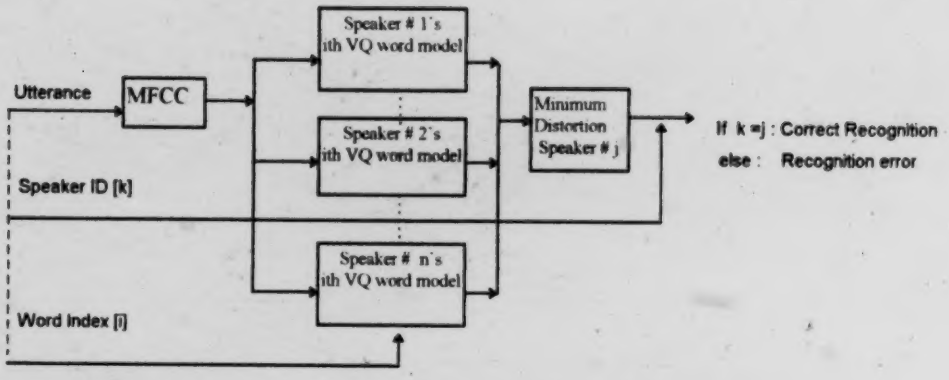

Utterance
MFCC
Speaker # 1's ith VQ word model
Speaker # 2's ith VQ word model
Speaker # n's ith VQ word model
Minimum Distortion Speaker # j
If k =j : Correct Recognition
else : Recognition error
Speaker ID [k]
Word Index [i]

Figure 6-10 Block diagram of the speaker identification error rate measurement.

The speaker identification test is designed in such a way that for each testing utterance, the program automatically identifies its word index and loads all the speakers' VQ codebooks with the same word index. Then, this utterance is matched with all those codebooks and finds the one with the least VQ distortion score. If the codebook has the same User ID as the utterance's, then, this utterance is counted as a correct identification. Otherwise, it is counted as an identification error. A speaker's identification error rate (IER) is the percentage of the identification errors.

For an estimation of speaker information coded in each MFCC coefficient, the baseline performs a speaker identification test for each speaker, with all the MFCC coefficients included in the VQ distortion measure. The subsequent tests repeat that same testing procedure except that each one of the 15 coefficients was excluded in turn from the VQ distortion measure. The normalized IER score (NIER), which is the difference score between the baseline IER score and the IER score that is calculated with one of the MFCC coefficients excluded, is used to indicate how the variance in that particular coefficient affects the speaker identification performance, or in other words, how much speaker information is coded in that MFCC coefficient.

Experimental Results

There are a total of 16 speaker identification tests, which include the baseline test and 15 tests in which each coefficient is excluded from the VQ distortion measure in turn. The results of individual speakers' NIER score distributions in the MFCC coefficients are listed in Appendix I-B. The corresponding graphs are presented in Appendices III-A and III-B. The average NIER distributions in the MFCC coefficients over all the speakers, the female speaker group and the male speaker group are plotted in Figures 6-11, 6-12 and 6-13. A positive NIER score indicates that the identification error rate increases with the exclusion of that MFCC coefficient. For example, in Figure 6-11, C_3 's NIER is 3.01%; this means that the exclusion of C_3 increases speaker identification error rate by 3.01%. A negative NIER score means that the exclusion of that MFCC coefficient decreases the speaker identification error rate by that amount. In other words, the variance contained in that MFCC coefficient is a confounding factor for speaker recognition, or that coefficient contains negative speaker information.

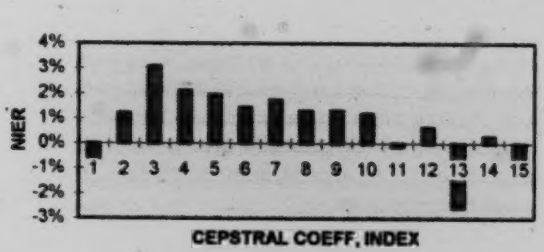


Figure 6-11 Average NIER score distribution in the MFCC over all the speakers.

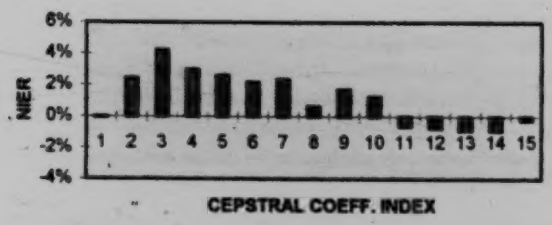


Figure 6-12 Average NIER score distribution in the MFCC over the female speaker group.

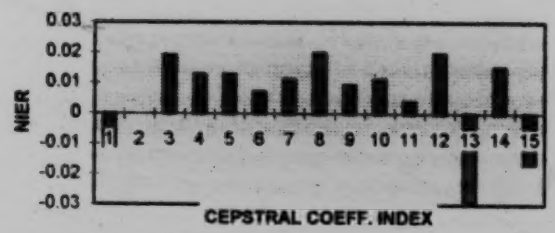


Figure 6-13 Average NIER score distribution in the MFCC over the male speaker group.

Comparison of the NIER score distribution in Figures 6-11, 6-12 and 6-13 with the corresponding NIDD score distribution in Figures 6-7, 6-8 and 6-9 shows that the two different speaker-information measurements yielded similar distribution patterns of speaker information in the cepstral domain. The Pearson product-moment correlation coefficient (Eq. 6.8) was computed to find out the correlation between the results of these two different methods :

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{[n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2][n\sum_{i=1}^{n} yi^2 - (\sum_{i=1}^{n} y_i)^2]}} \qquad (6.8)$$

where n is the number of cepstral coefficients, which in our experiment is 15; x_i is the ith cepstral coefficient's average NIDD score; y_i is the ith cepstral coefficient's average NIER score. The resulting coefficient of linear correlation is 0.796.

Though the IDD and SIER methods have yielded similar average distribution patterns of speaker information, there still exists a discrepancy as far as individual coefficients are concerned. It is especially true when we compare the individual speakers' NIDD and NIER score distributions (see Appendices I-A & I-B, Appendices II-A & II-B and Appendices III-A & III-B). For some speakers, the distribution patterns of speaker information in the MFCC coefficients based on the IDD and the SIER methods are quite different (e.g. , female speaker F2 and male speaker M4). An evaluation will be performed in the next chapter to see which method provides a better estimation of the speaker-information distribution in the MFCC coefficients.

6.9 Phonetic-Information Distribution

To verify our hypothesis that speaker- and phonetic-information distributions in the cepstral domain have their distinctive patterns, the phonetic-information distribution in the MFCC is also investigated. The phonetic information in each MFCC coefficient is estimated by using *the speech recognition error rate measurement* (SRER).

The difference between SRER and SIER is that in the SIER, the testing utterance matches with all speakers' same word models, and finds the model which produces the least distortion score. If the testing utterance's speaker ID matches with the model's speaker ID, it is counted as a correct speaker identification. Otherwise, it is counted as a speaker identification error. While in the SRER, the testing utterance matches with all the word models of the same speaker, and finds the one which produces the least distortion score. If the model's word index matches with that of the testing utterance's, then the utterance is correctly

recognized. Otherwise, it is counted as a speech recognition error. For compatibility with the speaker-information investigation, the SRER uses the same VQ approach for word modeling. Though the temporal information of the utterances will be lost with the VQ, it does not affect our investigation because what we are interested in is finding only the spectral information coded in the MFCC. The speech recognition procedure is illustrated in Figure 6-14 :

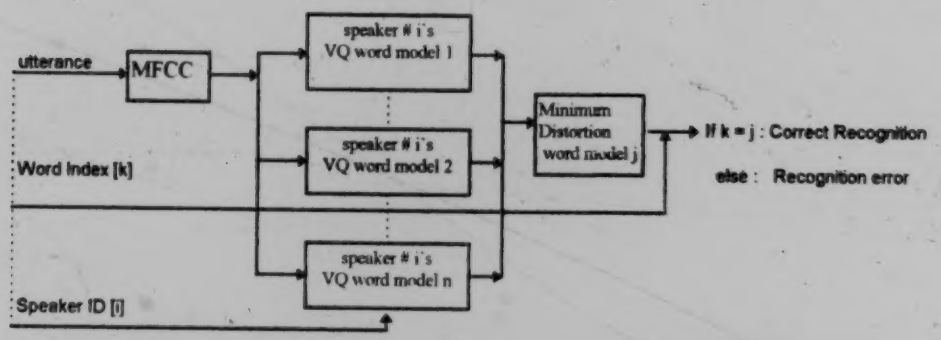


Figure 6-14 Block diagram of the speech recognition error rate measurement.

For the SRER measurement, the same TI-20 data are used. The word models used are built in exactly the same way as for the SIER measurement. The testing data are also the same. The testing procedure is designed in such a way that it loads utterances in sequence. With each utterance, the program automatically identifies its speaker ID according to the wave file name, loads all the word models of that speaker, calculates the VQ distortion scores, selects the closest codebook and matches its word index with the testing utterance's word index. For each speaker, the speech recognition errors are calculated and the phonetic information coded in each MFCC coefficient is approximated by the normalized speech recognition error rate (NRER). The NRER score is the difference between the baseline speech recognition error rate, which includes all the 15 MFCC

coefficients in the VQ distortion measurement, and the one which excludes one MFCC coefficient.

Results

Same as the NIER measurement, the SRER measurement is performed with the rotation of each of the 15 MFCC coefficients excluded from the VQ distortion measurement, as well as a baseline test which includes all the 15 MFCC coefficients. The resulting individual speakers' NRER score distributions in the MFCC are listed in Appendix IV. The graphs are presented in Appendices V-A and V-B. The average NRER score distributions in the MFCC coefficients over all the speakers, the female and the male speaker groups, are presented in Figures 6-15, 6-16 and 6-17.

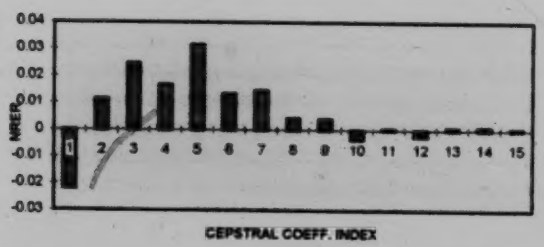


Figure 6-15 The average NRER score distribution in the MFCC over all the speakers.

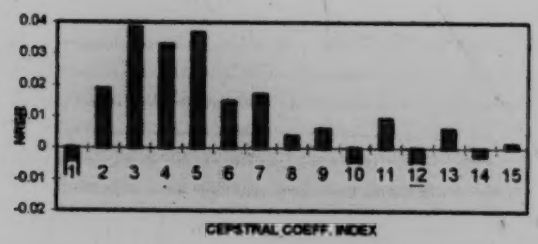


Figure 6-16 The average NRER score distribution in the MFCC over the female speaker group.

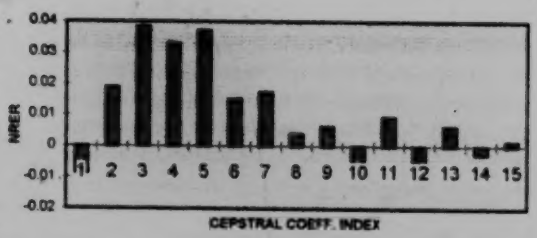


Figure 6-17 The average NRER score distributions in the MFCC over the male speaker group.

6. 10 Discussion

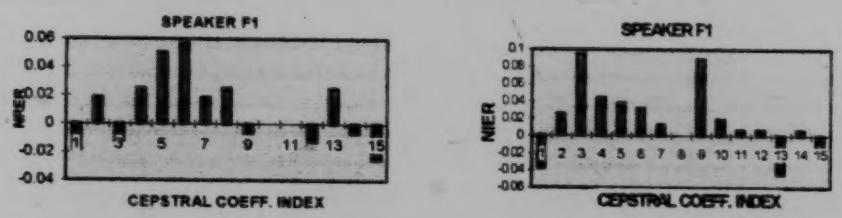
Figures 6-11, 6-12 and 6-13 show the average speaker-information distributions in the MFCC using the SIER measurement for the TI-20 data. Figures 6-15, 6-16 and 6-17 show the average phonetic-information distributions for the same data. There exist differences between the speaker- and phonetic-information distributions. However, both distributions also share some global similarity, that is, the lowest and higher-order cepstral coefficients contain less or little phonetic and speaker information. This global similarity can be partially due to the speech recording and processing environment. As pointed out by Juang and Rabiner (1987), the variability of the higher-order cepstral coefficients is partially caused

by the inherent artifacts of the signal processing procedure. This diminishes the discriminating power of the higher-order cepstral coefficients. The channel differences, such as the effect of differences in channel frequency response rolloff, on the other hand, usually affect most the first couple of cepstral coefficients and consequently degrade their performance.

The correlation between the average speaker- and phonetic-information distributions in the MFCC coefficients was computed. The resulting coefficient of linear correlation is 0.66. This suggests that the distributions of speaker and phonetic information have a fairly high correlation in the cepstral domain. It is consistent with the result reported by Furui (1986). Furui experimented with a linear model which consists of the phonetic factor, the speaker factor and the interaction between phonetic and speaker factors. The effects of phonetic, speaker and interaction factors were measured based on the multivariate analysis of variance using χ^2 distributions. His finding was that the phonetic effect was much larger than the speaker effect, and the interaction effect was also relatively large.

Since the general speaker- and phonetic-information distributions represent the average value from all the speakers in the database, they have obscured individual speakers' differences. The general speaker- and phonetic-information distribution patterns only provide us with the general tendency. For both speaker- and speech-information distributions, each individual speaker has his/her own unique patterns (see Appendices II-A & II-B, III-A & III-B, V-A & V-B), and they can be quite different from the average distribution patterns. For example, in the average distribution pattern of speaker information using the SIER measurement (Figure 6-11), the highest cepstral coefficient C_{15} is negative in speaker information. In other words, the variability contained in this coefficient is rather a confounding factor for speaker recognition. However, a look at Appendices III-A and III-B reveals that there are 6 speakers (36.5% of all speakers) who actually have positive speaker-information values for this coefficient.

The speaker-information distribution is mostly speaker-dependent. This is also true for the phonetic-information distribution. To compare the difference between the speaker- and phonetic-information distribution patterns, the best way is to look at individual speakers' patterns, instead of the averaged patterns. As an illustration, we present speaker F1's phonetic- and speaker-information distribution in the MFCC for a closer comparison.



A. The phonetic-information distribution B. The speaker -information distribution

Figure 6-18 Speaker F1's phonetic- and speaker-information distributions in the MFCC.

From Figure 6-16 we can see that the speaker- and phonetic-information distributions for speaker F1 are both similar and different. They are similar in the way that the lowest and higher-order coefficients in both distribution patterns have less relevant information. This conforms with the general distribution patterns for both kinds of information. The difference is that those coefficients which contain high speaker information tend to be low in phonetic information. This demonstrates, on the one hand, that speaker and phonetic information have their distinctive representations in the parametric domain. On the other hand, it suggests that significant speaker variances contained in those coefficients are confounding factor for speech recognition.

In summary, we have investigated the distribution patterns of speaker information in the MFCC. Two statistical methods (IDD and SIER) have been used independently for speaker-information measurement. They have yielded close

results concerning the distribution patterns of speaker information, which can be briefly summarized as follows:

- In general, the lowest- and the higher-order MFCC coefficients tend to contain less speaker information compared with the lower- and middle-order ones; for many speakers, the variances contained in the lowest- and higher-order MFCC coefficients are confounding factors for speaker recognition.
- The female speakers tend to have much less speaker information distributed in the higher-order MFCC coefficient region as compared with the male speakers.
- Though there exist the above general tendencies in the distribution patterns of speaker information, the amount of speaker information in each MFCC coefficient is largely speaker-dependent. In other words, each individual speaker's idiosyncrasy is coded in the MFCC in its unique way.

In this chapter we have also investigated phonetic-information distribution in the MFCC. Comparison of the distribution patterns of phonetic and speaker information shows that these two kinds of information have their distinct distribution patterns in the cepstral domain. However, there also exists fairly high correlation between these two distributions.

CHAPTER 7
SPEAKER-INFORMATION ENHANCEMENT

In the last chapter, we have investigated speaker-information distribution patterns in the MFCC coefficients. Based on those findings, we will compare different weighting strategies in the cepstral domain in search of an optimal way for speaker-information enhancement.

7.1 Experimental Design

A baseline and three different cepstrum weighting (or "liftering", a special name in cepstral analysis, analogous to filtering) methods are implemented in the closed-set speaker identification tests for comparison. The baseline speaker identification test uses only the inverse-variance weighting (see Chapter 6: 6.6). For the other three methods, different weighting functions are applied, in addition to the inverse-variance weighting. ***Weighting Function A*** uses the raised sine function (Juang et al. , 1987), a popularly used cepstral weighting strategy for speech information enhancement; ***Weighting Function B*** uses the average distribution pattern of speaker information in the MFCC acquired from the last experiment; ***Weighting Function C*** uses the individual speakers' distribution patterns of speaker information in the MFCC. For comparison of the IDD and SIER methods, Weighting Functions B and C include two sub-experiments which use distribution patterns of speaker information provided by the IDD and SIER measurements in the last experiment respectively. The above three weighting methods can also be classified into two categories: *the general weighting* approach and *the speaker-based weighting* approach. For *the general weighting* approach, the same weighting function is applied to all the speakers indiscriminately in the VQ Euclidean distance measurement.

$$d(x,x') = \sum_{i=1}^{p} (w_i |xi - x'i|^2 \, Mi). \qquad (7.1)$$

where *wi* is the *ith* MFCC coefficient's inverse-variance weight. M_i is the *ith* MFCC coefficient's speaker-information weight. The function of M_i is to enhance speaker information in that particular MFCC coefficient. In the *general weighting* approach, M_i is the same for all the speakers. *The general weighting* approach includes Weighting Functions A, B1 and B2.

For *the speaker-based weighting* approach, weighting is based on individual speakers' information distribution patterns in the MFCC and different weightings are applied to different speakers in the Euclidean distance measurement.

$$d(x,x') = \sum_{i=1}^{p} (w_i |xi - x'i|^2 \, M_{ji}). \qquad (7.2)$$

where M_{ji} is the *jth* speaker's *ith* cepstral coefficient speaker-information weight. Weighting Functions C1 and C2 are *the speaker-based weighting* approach.

Weighting Function A

This weighting function is a raised sine function (Figure 7-1), which was originally proposed by Juang et al. (1987) in their speech recognition experiment:

$$M(i) = 1 + 0.5p \sin(\pi i / p), \qquad 1 \le i \le p. \qquad (7.3)$$

where *p* is the order of cepstral coefficients. The purpose of this weighting function is to minimize the cepstral variability caused by the artefacts of the analysis procedure or sources which did not pertain to speech recognition.

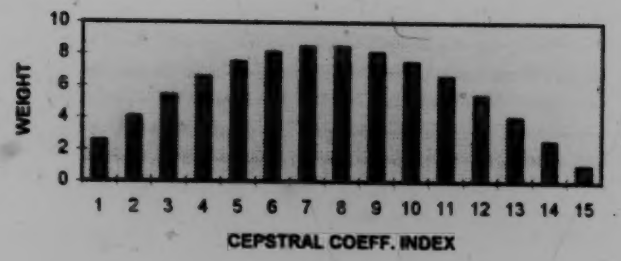


Figure 7-1 Raised sine weighting function.

As we have discussed in the previous chapter, the variability of the higher-order cepstral coefficients is partially caused by the inherent artefacts of the signal processing procedure; and channel differences, such as the effect of differences in the channel frequency response rolloff, usually affect the first couple of cepstral coefficients. Speaker variability also significantly affects the lower cepstral coefficients. Juang et al. (1987) tested different cepstral weighting methods to reduce the lower- and higher-order cepstral coefficients' effect on the distance measure and found that the raised sine function yielded a better speaker-independent speech recognition performance. We apply this raised sine function weighting for speaker identification in the present experiment. The purpose is to compare this conventional speech-information weighting approach with our new speaker-information weighting approaches.

Weighting Function B1

Weighting Function A was originally intended for speech recognition. Therefore, the lower-order coefficients are given low weighting because the variability of these coefficients is more related to speaker variation, as well as to the transmission channel variation. For speaker recognition, however, the variability due to speaker variation should be enhanced instead. Furthermore, the average distribution pattern of speaker information (see Figure 7-11) indicates that the speaker information distribution is actually not exactly a sine function and the

higher-order coefficients tend to have little or even negative speaker information. A better approach for speaker information enhancement, then, should adopt a strategy which weights the MFCC coefficients according to the average distribution pattern of speaker information, which is obtained from the speech training data. Weighting Function B1 is based on this approach, and weights each MFCC coefficient according to its ranking in the amount of speaker information. The coefficients are sorted in an increasing order according to their respective NIER scores based on the SIER measure. The weights rank from 1 to 15. The lowest-ranking coefficient is assigned the weight value of 1 and there is an increment of 1 for each subsequent coefficient. If two or more coefficients have the same NIER scores, they receive the same weight. In that case, the highest weight will be lower than 15. The general-speaker-information based weighting function is illustrated in Figure 7-2. This weighting function is applied to all the speakers in the testing phase.

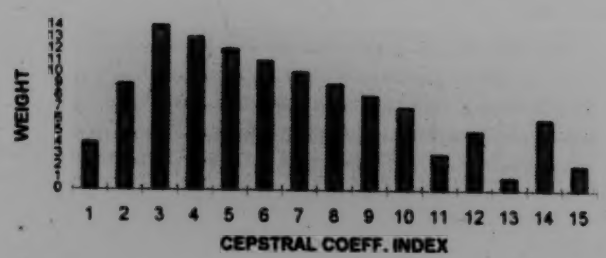


Figure 7-2 Weighting function based on the average NIER score distribution in the MFCC coefficients.

Weighting Function B2

For measurement of the speaker-information distribution, two different statistical methods were used in the last experiment: one is the IDD measurement; the other is the SIER measurement. Weighting Function B1 is based on the average

NIER score distribution in the MFCC using the SIER measurement. Weighting Function B2 uses the average NIDD scores based on the IDD measurement. The purpose of performing Weighting Function B2 is to find out which statistical method provided a better estimation of the average distribution pattern of speaker information in the cepstral domain. Weighting Function B2 is plotted in Figure 7-3.

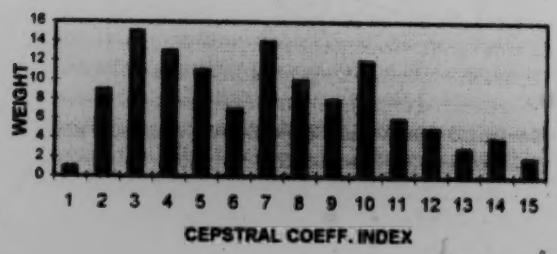


Figure 7-3 Weighting function based on the average NIDD score distribution in the cepstral coefficients.

Weighting Function C1

The difference between Weighting Functions B1 and B2 is that Weighting Function B1 is based on the average NIER score distribution in the MFCC coefficients, while Weighting Function B2 is based on the average NIDD score distribution. However, both weighting functions assume that the average distribution pattern of speaker information is applicable to all the speakers. Therefore, the same weighting function is applied to speakers indiscriminately in the testing phase.

As we have already pointed out, although there exists a general tendency, the speaker-information distribution is largely speaker-dependent. If we look at the individual speakers' NIDD or NIER score distribution patterns in the cepstral domain (Appendices II-A_B, III-A_B), we can see that there exist significant inter-speaker differences. Some individual speakers' patterns are quite inconsistent with the average distribution pattern of speaker information. For example, the

lowest- and higher-order coefficients contain little speaker information in the average distribution pattern, but this is not the case in some individual speakers' patterns. The middle-order coefficients contain much speaker information in the average distribution pattern. However, some individual speakers' patterns show negative NIER or NIDD scores in that region. Optimal speaker-information enhancement, then, has to depend on individual speakers' distribution patterns of speaker information.

Weighting Function C1 adopts this approach. The weight assigned to each coefficient is based on the ranking of NIER scores of the SIER measure. In this respect, Weighting Function C1 is the same as Weighting Function B1. However, the ranking of speaker information in Weighting Function B1 is based on the average NIER score distribution, and the weighting assigned to each coefficient is the same for all the speakers. The ranking of speaker information in Weighting Function C1, on the other hand, is based on each individual speaker's NIER score distribution and the weighting assigned to each coefficient is speaker-dependent (see Table 7-1).

Table 7-1: Cepstral coefficient weighting for individual speakers based on the NIER score

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15
F1	2	8	13	11	10	9	6	4	12	7	5	5	1	5	3
F2	1	1	1	2	3	3	2	1	1	2	2	2	3	1	4
F3	3	7	7	5	2	7	4	6	4	4	2	1	1	1	2
F4	8	9	10	8	5	6	6	4	6	7	1	6	2	4	3
F5	3	7	6	9	6	4	8	5	10	4	5	3	2	1	4
F6	7	9	10	7	10	11	9	6	2	4	3	4	1	5	8
F7	4	6	9	8	8	6	9	4	5	7	5	2	1	5	3
F8	4	3	9	8	10	4	7	5	3	6	1	2	3	2	2
M1	5	1	6	9	8	8	12	4	2	10	7	10	4	11	3
M2	2	4	6	9	8	8	6	7	9	4	3	5	4	5	1
M3	1	7	5	3	3	6	3	8	4	9	6	3	5	4	2
M4	3	3	8	7	5	4	5	6	2	3	1	2	2	3	3
M5	6	8	9	4	3	3	2	4	4	3	6	7	1	5	2
M6	1	2	9	5	12	4	6	11	10	5	8	7	3	11	8
M7	6	9	6	4	5	5	6	7	6	5	5	8	1	2	3
M8	10	5	4	7	2	4	3	6	8	5	6	9	1	8	5

Weighting Function C2

Weighting Function C1 applies different weighting patterns to different speakers. Weighting Function C2 adopts the same strategy, except that the weighting patterns are based on individual speakers' IDD score distributions instead of the NIER score distributions (see Table 7-2). The purpose of performing Weighting Function C2 is the same as for Weighting Function B2, which is to compare the performance of the two different statistical methods for speaker information estimation in the cepstral domain.

Table 7-2: Cepstral coefficient weighting for individual speakers based on the NIDD score

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15
F1	1	8	12	10	15	13	9	11	14	5	4	7	3	6	2
F2	14	15	9	12	10	8	2	13	7	11	6	3	4	5	1
F3	3	12	14	9	13	11	15	6	10	7	8	5	2	4	1
F4	5	14	15	13	9	11	8	4	10	12	7	6	3	2	1
F5	1	13	15	14	7	12	8	10	9	11	5	4	2	3	6
F6	1	12	15	10	13	6	14	2	11	9	8	4	5	7	3
F7	2	14	15	13	12	3	9	10	11	4	5	1	8	7	6
F8	1	6	15	10	8	3	14	13	11	7	12	4	9	2	5
M1	1	2	10	15	13	11	14	12	6	4	8	7	9	5	3
M2	1	14	12	8	9	13	10	7	15	11	3	4	5	6	2
M3	4	12	6	7	10	8	11	15	1	14	2	3	5	9	13
M4	10	2	3	4	7	15	8	14	6	11	5	9	12	13	1
M5	1	15	13	14	9	4	11	8	2	12	10	7	6	3	5
M6	1	6	15	12	10	4	9	13	11	14	3	7	2	8	5
M7	1	10	9	3	7	8	11	12	15	14	13	6	2	4	5
M8	2	6	7	15	11	12	10	14	5	9	4	13	3	8	1

For the baseline and all the weighting approaches, the text-dependent speaker models are trained on the recording session 1 of the TI-20 data. The testing data are from recording sessions 5 to 9, which consist of 8 repetitions of the same vocabulary item for each speaker. In the testing phase, the MFCC coefficients of each frame of the utterance were weighted according to different weighting approaches in the Euclidean distance measure for the VQ distortion score. The identification error rate (IER) for each speaker is calculated the same way as in the SIER measurement (Chapter 6: 6.8).

7.2 Experimental Results

The IERs for the 16 speakers of the TI-20 data with the baseline and three different weighting approach tests are listed in Table 7.3. F1-8 are female speakers and M1-8 are male speakers.

Table 7-3: Speaker identification error rates for different weighting strategies

SPEAKER IDENTIFICATION TESTS						
Speaker	Baseline	Weighting A	Weighting B1	Weighting B2	Weighting C1	Weighting C2
F1	39.38%	29.38%	27.50%	31.88%	27.75%	29.38%
F2	5.00%	4.38%	5.63%	3.13%	2.50%	4.38%
F3	44.38%	32.50%	31.88%	34.38%	30.00%	30.63%
F4	31.25%	33.13%	31.25%	21.25%	26.88%	31.25%
F5	54.38%	47.75%	43.13%	39.38%	45.63%	42.50%
F6	39.38%	41.25%	27.13%	37.13%	31.25%	31.25%
F7	16.25%	11.88%	5.63%	11.88%	9.38%	6.88%
F8	27.13%	27.13%	21.25%	32.50%	23.13%	23.13%
M1	51.28%	46.79%	40.38%	33.97%	41.67%	41.03%
M2	42.50%	36.25%	37.50%	34.38%	35.63%	37.13%
M3	1.88%	1.88%	1.88%	4.38%	1.88%	1.88%
M4	26.28%	27.56%	24.36%	26.92%	22.44%	24.36%
M5	37.46%	36.54%	32.69%	29.49%	33.97%	33.33%
M6	32.08%	22.01%	23.90%	19.50%	20.75%	21.38%
M7	46.54%	43.40%	39.62%	47.17%	41.51%	41.51%
M8	43.75%	41.88%	43.75%	37.50%	31.88%	47.50%
Average	**33.81%**	**30.36%**	**27.40%**	**27.86%**	**26.70%**	**27.03%**

7.3 Discussion

The results of the three different weighting functions for speaker-information enhancement provide some interesting observations for discussion. Weighting Function A reduced the overall error rate by 3.45% in speaker identification performance compared with the baseline. Since this weighting function was originally designed for enhancing phonetic information only, the effect of this function in improving speaker recognition supports our experimental result in Chapter 6 that there exists a fairly strong correlation between the distribution patterns of speaker and phonetic information in the MFCC. One source contributing to the correlation between the distributions of phonetic and speaker information is from the speech processing environment, which mostly affects the lowest- and higher-order coefficients. The other source is from the acoustic

nature of speech itself. As pointed out by O'Shaughnessy (1987): "Most of the parameters and features used in speech analysis contain information useful for the identification of both the speaker and the spoken message. (p.480)". In the speech signal, the same acoustic phenomena, such as formant frequencies, carry both phonetic and speaker cues. Recent studies on speech perception (Goldinger et al. , 1991; Palmeri et al. , 1993; Nygaard et al. , 1994) further indicate that speaker information actually facilitates listeners' phonetic processing for some perceptual tasks. This suggests that the human's perceptual system treats speaker information as an integrated component of the acoustic cues for speech recognition, and there is an inherent relationship existing between phonetic and speaker information.

In spite of the fact that the implementation of Weighting Function A improved speaker recognition performance in general, there were three speakers (Speaker F4, F6 and M4) whose error rates actually were increased and another two speakers (F8 and M3) whose error rates remained the same, compared with the baseline (see Figure 7-4).

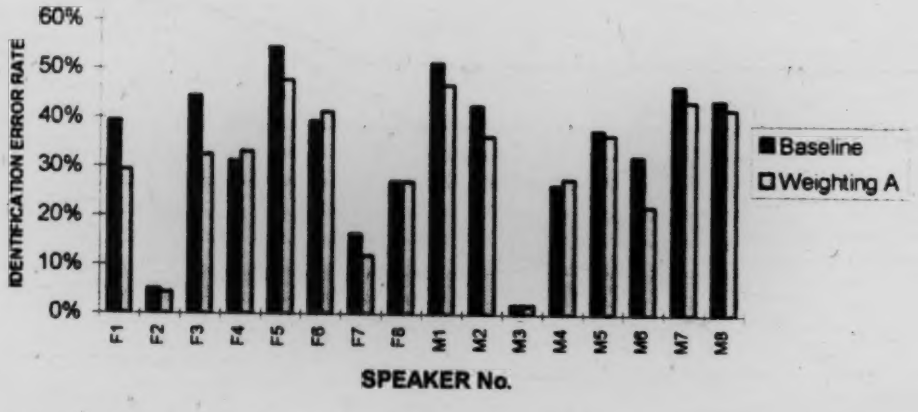


Figure 7-4 Comparison of speaker identification performance between the baseline and weighting function A.

Weighting Functions B1 and B2 were based on the average distribution pattern of speaker information in the MFCC according to either NIER or NIDD scores and they both achieved overall better performance than the baseline with significant error rate reduction of 6.41% and 5.95% respectively. They also outperformed Weighting Function A with error rate reduction of 2.84% and 2.50% respectively. These results strongly support our basic argument that speaker information has its distinct distribution pattern in the acoustic and parametric domain, which can be identified and enhanced effectively for speaker-recognition.

A further comparison between Weighting Function B1 and B2 shows that Function B1 achieved slightly better overall performance than Function B2. What is significant, however, is that for Weighting Function B1, there was only one speaker (F2) whose error rate actually deteriorated and 3 speakers (F4, M3 and M8) whose error rates remained the same, while all the other speakers improved their speaker identification performance. For Weighting Function B2, there were four speakers (F8, M3, M4, M7) whose error rate actually increased compared with the baseline (see Figure7.5).

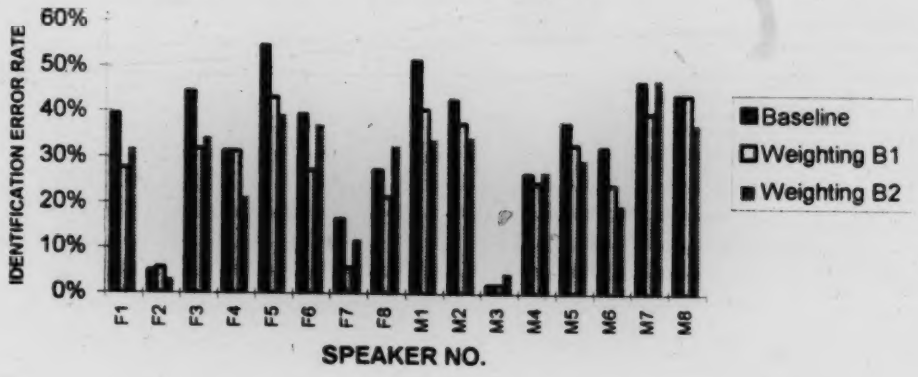


Figure 7-5 Comparison of speaker identification performances among the baseline, Weighting Function B1 and B2.

The better performance of Weighting Function B1 over Weighting Function B2 suggests that the SIER provided a better average speaker-information-distribution estimation than the IDD measurement.

Weighting Functions C1 and C2 were based on the individual speakers' NIER and NIDD score distributions respectively. In Comparison with the baseline and Weighting Function A, C1 and C2 performed significantly better. They also yielded slightly overall better performance than Weighting Functions B1 and B2. One advantage of C1 and C2 over B1 and B2, however, is that this individual-speaker-information-distribution-based approach reduced most of the individual speakers' identification error rates, rather than just the average error rate over all the speakers. In other words, most of the speakers were benefited from this weighting approach. It is particularly true for C1. Compared with the baseline, C1 reduced all individual speakers' identification error rate except one speaker (M3), whose error rate remained the same (see Figure 7-6). As for this particular speaker, there may be an explanation for the lack of improvement even with the use of speaker-dependent weighting strategy. The error rate for this particular speaker in the baseline is the lowest (1.88%). Compared with the average error rate 33.81%, improvement in performance for this speaker might have already been saturated.

C2 produced less satisfactory results than C1. There is still one speaker (M8) whose error rate actually increased and two speakers (F4 and M3) whose error rates did not change as compared with the baseline. This further indicates that the SIER provided better estimation of the speaker-information distribution than the IDD measurement.

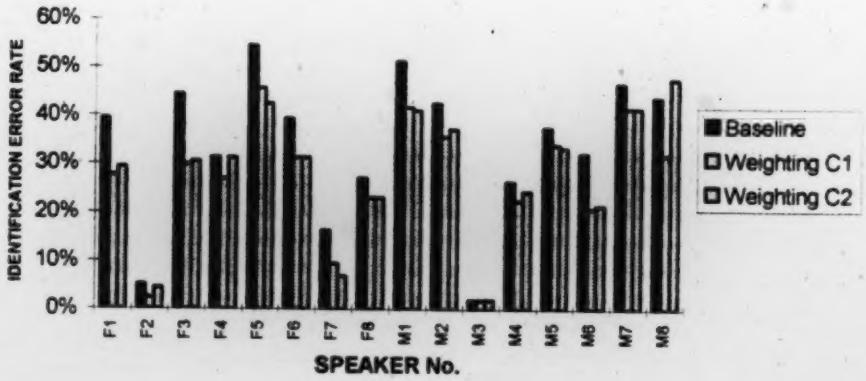


Figure 7-6 Comparison of the speaker identification performances among the baseline, Weighting Function C1 and C2.

Based on the above experimental results, we tentatively conclude that::

- Weighting based on the speaker-information distribution (B1, B2, C1 and C2) performs better than the conventional speech weighting method (Weighting Function A) for speaker-information enhancement.
- Weighting based on individual speakers' speaker-information distribution (C1 and C2) has one important advantage over the weighting approach based on the average speaker-information distribution (B1 and B2), that is, it is basically effective for all the speakers.

The better speaker identification performance with using individual speakers' distribution patterns conforms with a voice perception theory that different acoustic cues are used in distinguishing different voices (Lancker et al. , 1985). According to Lancker et al. , the critical parameter(s) for speaker information are not the same for all voices. The acoustic cue(s) essential for distinguishing one speaker's voice may be expendable in the case of distinguishing another speaker's voice. Loss of certain parameter(s) will not impair recognisability if a

voice is sufficiently distinctive on some other dimensions. In their voice perception experiments, Lancker et al. found that speech contains a constellation of potential cues from which the listener "selects" a subset to use for identifying a given voice. The weighting approach based on the individual-speaker-information distribution conforms with this basic human perceptual process.

As the two speaker-information-estimation methods are concerned, the SIER measure provides better speaker-information-distribution estimation than the IDD method. Compared with the IDD method, the SIER has the obvious advantage that its measurement comes directly from the speaker identification performance. The disadvantage of the SIER, however, is that this measure is not as sensitive as the IDD method in detecting the effect of a slight experimental condition changes. It is especially true in the situation when a speaker's intra-speaker and inter-speaker distance score distributions do not overlap with each other. In that case, the perturbation of the intra-speaker and inter-speaker distance score distributions will be reflected in the IDD score, but not in the IER score.

CHAPTER 8
SUMMARY AND CONCLUSION

The basic assumption in this study is that speech contains both phonetic (linguistic) and speaker information. Acoustically, these two kinds of information have their distinctive representations in the speech signal. The optimal approach for improving speaker recognition performance is to enhance only the speaker-information component coded in the speech signal.

This study first investigated the articulatory and acoustic aspects of speaker information, the interrelationship between speaker information and its phonetic environment, and the contrast between speaker and phonetic information cues. On the basis of speaker-information analysis, we proposed a new approach for speaker information enhancement. This approach used the speech training data to identify the speaker-information distribution in the parametric domain. In the testing phase, corresponding weighting (or liftering) strategy was applied to enhance the speaker-information rich elements. Since the cepstrum is a widely used parametric representation of the speech signal in both speech and speaker recognition systems, this study was focused in the methods for measurement of speaker information in the MFCC, and the optimal weighting strategy for speaker-information enhancement.

The first part of the experiments was to measure the speaker-information distribution in the MFCC. Two statistical methods were used independently for speaker-information estimation. One was the IDD method, which measured the intra- and inter-speaker distribution scores. The other was the SIER method, which measured the speaker identification error rates. In both methods, Session 1 of the TI-20 speech data was used for speaker VQ model training, and Sessions 2-5 were used for speaker-information measurement. The experimental results from both statistical methods showed similar general distribution patterns of speaker information in the MFCC coefficients. In general, the lowest- and higher-order coefficients contained little speaker information, or even confounding

variances; most of the speaker information was concentrated in the lower- and middle-order coefficients. As for the sex difference, female speakers tended to have much less speaker information distributed in the region of the higher-order coefficients, as compared with male speakers. In spite of the above general tendencies, the speaker-information distribution in the MFCC was found to be largely speaker-dependent. There were cases in which individual speakers' information distribution patterns did not conform with the general distribution pattern of speaker information or the distribution pattern of its own sex group. The phonetic-information distribution pattern in the cepstral domain was also investigated to verify our assumption that speaker information and phonetic information have their distinctive representations. Comparison of the general phonetic- and speaker-information distributions in the MFCC indicated that the two information distribution patterns were different. However, they also had relatively-high correlation ($r = 0.66$). This correlation could come from two sources: one source was the inherent overlap between the phonetic and speaker information. The other source was the channel effect and artefacts of the signal processing procedure. Further comparison of individual speakers' phonetic- and speaker-information distribution patterns showed that the MFCC coefficients which contained high speaker information tended to be low and even negative in phonetic information. This suggests that speaker variances can be serious confounding factors in speech recognition.

The second part of the experiment was in search of the optimal speaker-information enhancement strategy in the cepstral domain. We compared three different weighting functions for speaker-information enhancement. Weighting Function A was a raised sine function, which was originally proposed by Juang at el. for speech enhancement, not specifically targeted for speaker information. Weighting Functions B (1-2) were based on the average distribution pattern of speaker information in the MFCC coefficients. The weight assigned to each coefficient was proportional to either its average NIER or NIDD score. Weighting Functions A, B (1-2) applied the same weighting function to all the speakers

indiscriminately. Different from Weighting Functions A and B(1-2), Weighting Functions C (1-2) were based on the individual speakers' NIER or NIDD score distribution patterns in the MFCC coefficients. For different speakers, Weighting Function C (1-2) applied different weightings.

The three weighting functions were tested in the VQ-based text-dependent speaker identification program. Sessions 6-9 of the TI20 data were used as testing stimuli. The experimental results show that all three weighting functions improved speaker recognition performance compared with the baseline; however, the weighting functions based on speaker information performed better than the weighting function based on speech information. Furthermore, the weighting Function which was based on individual speaker information gave an overall better speaker identification performance. This result supported a voice perception theory, which suggested that a speaker's voice pattern contains a constellation of potential cues from which the listener "selects" a subset to use for identifying a given voice.

With Weighting Functions B and C, we also compared two sets of speaker-information scores (NIER and NIDD), which resulted from two different statistical methods for speaker-information measurement. The speaker identification results indicated that the SIER measurement produced an overall better speaker-information estimation than the IDD measurement.

In conclusion, this study has developed ideas on how to improve automatic speaker recognition via a new methodology and technique designed to enhance those elements in the speech parameters most relevant to discriminate speakers. By increasing the weights in appropriate distance measures along the lines of the power of each parameter to discriminate speakers, speaker information is enhanced, which improves the performance of speaker recognition. The same methodology can be applied in speech recognition to enhance the phonetic information. How well this approach will improve speech recognition performance is our next research interest.

BIBLIOGRAPHY

Ainsworth, W. A. (1976). Mechanisms of Speech Recognition. Oxford: Pergamon Press.

Andrews, G. , Platt, L. J. , & Young, M. (1977). Factors affecting the intelligibility of cerebral palsy speech to the average listener. Folia Phoniatrica 29, 292-301.

Ariki, Y. , & Doi, K. (1994). Speaker recognition based on subspace methods. Proc. ICSLP, S31-7.1, 1859-1862.

Assaleh, K. T. & Mammone, R. J. (1994). Robust cepstral features for speaker identification. Proc. ICASSP-94, I-129-132.

Atal, B. S. (1972). Automatic speaker recognition based on pitch contours. J. Acoust. Soc. A.m. 52, 1687-1697.

Atal, B. S. (1974). Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. J. Acoust. Soc. Am. 55,1304-1312.

Berg, J. van den (1958). Myoelastic-aerodynamic theory of voice production. Journal of Speech and Hearing Research I, 227-244.

Bernasconi, C. (1990). On the instantaneous and transitional spectral information for text-dependent speaker verification. Speech Communication 9, North Holland , 129 - 139.

Bladon, R. A. , & Lindbbm, B. (1981). Modeling the Judgment of vowel quality differences. J. Acoust. Soc. Am. 69, 1414-1422.

Bonastre, J. F. , Meloni, H. , & Langlais, P. (1991). Analytical strategy for speaker identification. Proc. EUROSPEECH-91 2, 435-438.

Borden, G. J. , & Harris, K. S. (1980). Speech Science Primer. London: Williams & Wilkins.

Bordone-Sacerdote, C. & Sacerdote, G. G. (1969). Some spectral properties of individual speakers. Acoustica 21, 199-210.

Bricker, P. D. , & Pruzansky, S. (1966). Effects of stimulus content and duration on talker identification. J. Acoust. Soc. Am. 40, 1,441-1,449.

Bruyninckx, M. , Harmegnies, B. , Llisterri, J. , & Poch-Olive, D. (1994). Language-induced voice quality variability in bilinguals. Journal of Phonetics 22, 19-31.

Chen, F. Macleod, I. , Millar, B. , & Laverty, W. (1994). Optimal cohort design in VQ-distortion based text-independent speaker verification. Proceedings of the Fifth Australian International Conference on Speech Science and Technology, 750-755, Perth, Australia.

Chen, F. , & Roszypal, A. (1991). Computer modeling of lexical tone perception. Canadian Acoustics 19, 103-104.

Chen, F. , & Roszypal, A. (1992), Computer Mandarin Tone perception. ICA Proceedings 3, G2-6, Beijing, China.

Chen, F. , Millar, B. , & Wagner, M. (1994). Hybrid threshold approach in text-independent speaker verification. Proc. ICSLP, 1855-1858.

Chengalvarayan, R. , & Deng, L. (1997). Use of generalized dynamic feature parameters of speech recognition. IEEE Trans. On Speech and Audio Processing 5, No. 3, 232-242.

Clements, G. N. (1985). The geometry of phonological features. Phonology Yearbook 2, 223-225.

Craner, B. , & Schroeter, J. (1995). Modeling a leaky glottis. Journal of Phonetics 23, 165-177.

Das, S. K. , & Mohn, W. S. (1971). A scheme for speech processing automatic speaker verification. IEEE Trans. Audio Electro-acoust. AU-19, 32-43.

Davis, S. B. & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. on ASSP 28, No.4, 357-366.

Delattre, P. C. (1951). The physiological interpretation of sound spectrograms. PMLA, 66, 864-875.

Deller, J. R. , Proakis, J. G. , & Hansen, J. H. L. (1993). Discrete-time processing of speech signals. New York: Macmillan Publishing Company.

Diehl, C.F. , White, F. , & Burk, K. W. (1959), Voice quality and anxiety. Journal of Speech and Hearing Research 2, 282-285.

Diehl, R. L. , Lindblom, B. , Hoemeke, K. A & Fahey, R. P. (1996). On explaining certain male-female differences in the phonetic realization of vowel categories. Journal of Phonetics 24, 187-208.

Doddington, G. R. (1971). A new method of speaker verification. J. Acoust. Soc. Am. 49, 139(A).

Eatock, J. P. , & Mason, J. S. (1994). A quantitative assessment of the relative speaker discriminating properties of phonemes. Proc. ICASSP 94, I-133-136.

Elyan, O. (1978). Sex differences in speech style. Women Speaking, 4. April.

Endres, W. , Bambach, W. , & Flosser, G. (1971) Voice spectrograms as a function of age, voice disguise and voice imitation. J. Acoust. Soc. Am. 49, 1842-1848.

Fakotakis, N. , Tsopanoglou, A. , & Kokkinakis, G. (1991). Text independent speaker recognition based on vowel spotting. Proc. IEE 6th Internat. Conf. Processing of Signals in Communications, Loughborough, 2-5.

Fakotakis, N. , Tsopanoglou, A. , & Kokkinakis, G. (1993). A text-independent speaker recognition system based on vowel spotting. Speech Communication 12, 57-68.

Fant, G. (1960). Acoustic Theory of Speech Production. The Hague: Mouton.

Fant, G. (1968). Analysis and synthesis of speech processes. In B. Malmberg (Ed.), Manual of Phonetics, Amsterdam.

Fant, G. (1973). Speech Sounds and Features. Cambridge: The MIT Press.

Floch, J-L. L. , Montacie, C. , & Caraty, M-J. (1994). Investigations on speaker characterization from Orphee system techniques. Proc. ICASSP-94, I-149-152.

Frokjaer-Jensen, B. , & Prytz, S. (1976). Registration of voice quality, Bruel and Kjaer Technical Review 3, 3-17.

Fujimura, O. (1962). Analysis of Nasal Consonants. J. Acoust. Soc. Am. 34, 1865-1875.

Fujimura, O. , & Lindqvist, J. (1971). Sweep-tone measurements of vocal tract characteristics. J. Acoust. Soc. Am. 49, 541-558.

Furui, S. (1981). Cepstral analysis technique for automatic speaker verification. IEEE Trans. on ASSP 29, 254-272.

Furui, S. (1986). Speaker independent isolated word recognition using dynamic features of speech spectrum. IEEE Trans. on ASSP, 34, 52-59.

Furui, S. (1986). Research on individuality features in speech waves and automatic speaker recognition techniques. Speech Communication 5, 183-197.

Furui, S. (1989). Digital Speech Processing, Synthesis, and Recognition. New York and Basel: Marcel Dekker, Inc. .

Furui, S. (1991). Speaker-dependent-feature extraction, recognition and processing techniques. Speech Communication 10, 505-520.

Furui, S. (1994). An overview of speaker recognition technology. ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, Martigny, Switzerland, 1-9.

Furui, S. , Itakura, F. , & Saito, S. (1972). Talk recognition by longtime averaged speech spectrum. Trans. IECE, 55-A, 549-556.

Garvin, P. , & Ladefoged, P. (1963). Speaker identification and message identification in speech recognition. Phonetica, 9, 193-199.

Gerstman, L. (1957). Cues for distinguishing among fricatives, affricates and stop consonants. Unpublished doctoral dissertation, New York University.

Glenn, J. K. , & Kleiner, N. (1968). Speaker identification based on nasal phonation. J. Acoust. Soc. Am. 43, 368-372.

Glodinger, S. D. , Pisoni, D. B. , & Logan, J. S. (1991). On the nature of talker variability effects on recall of spoken word lists. J. Exp. Psychol. 17. 152-162.

Goldstein, U. G. (1976). Speaker-identifying features based on formant tracks. J. Acoust. Soc. Am. 59, 176-182.

Hargreaves, W. , & Starkweather, J. A. (1963). Recognition of speaker identity. Language and Speech 6, 63-67.

Hayakawa, S. , & Itakura, F. (1994). Text-dependent speaker recognition using the information in the higher frequency band. Proc. ICASSP-94, I-137-140.

Herbst, L. (1969). Die Umfange der physiologischen Hauptsprechtonbereiche von Frauen und Mannern. Zeitschrift fur Phonetic. 22, 426-438.

Hermansky, H. & Junqua, J. C. (1988). Optimization of perceptually-based processing ASR front-end. Proc. ICASSP, 219-222.

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. J. Acoust. Soc. Am. 87, 1738-1752.

Hermansky, H. , Tsuga, K. , Makino, S. , & Wakita, H. (1985). Perceptually based processing in automatic speech recognition. Proc. ICASSP, 1971-1974.

Heuvel, H. , & Rietveld, T. (1992). Speaker related variability in cepstral representations of Dutch speech segments. Proc. ICSLP 2, 1581-1584.

Heuvel, H. , Cranen, & Rietveld, T. (1996). Speaker variability in the coarticulation of / a, i, u /. Speech Communication 18, 113-130.

Hofker, U. (1977). AUROS-Automatic recognition of speakers by computer, phoneme-ordering for speaker recognition. 9th International Congress on Acoustics, Madrid, 506.

Hogan, J. T. (1996). A study of Kono tone spacing. Phonetica 53, 221-229.

Hoit, J. , & Hixon, T. (1987). Age and speech breathing. Journal of Speech and Hearing Research 30, 351-366.

Homayounpoiur, M. M. , Chollet, G. , Goldman, J. , & Vaissiere, J. (1993). Performance comparison of machine and human speaker verification. SPEUROEECH 93, 2295-2298.

Homayounpour, M. M. , & Chollet, G. (1994). A comparison of some relevant parametric representations for speaker verification. Workshop on Automatic Speaker Recognition, Martigny, Switzerland, pp. 185-188.

Inbar, G. F. , & Eden, G. (1976). Psychological stress evaluators: EMG correlation with voice tremor. Bilolgical Cybernetics 24, 165-167.

Jack, M. , & Laver, J. (1988). Aspects of Speech Technology. Edinburgh: University Press.

Jakobson, R. , Fant, G. , & Halle, M. (1952). Preliminaries to speech analysis. The distinctive features and their correlates. Acoustics Laboratory, Technical Report No.13, MIT.

Juang, B-H. , Rabiner, L. R. , & Wilpon, J. G. (1986). On the use of bandpass liftering in speech recognition, IEEE Trans. on ASSP 35, 947-954.

Kao, Y. , Baras, J. , & Rqjasekaran, P. (1993). Robustness study of free-text speaker identification & verification. Proc. ICASSP, 379-382.

Klatt, D. H. , & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. J. Acoust. Soc. Am. 87, 820-857.

Kuwabara, H. , & Oogushi, K. (1983). Acoustic characteristics of professional male announcers' speech. Trans. IECE, J66-A, 545-552.

Ladefoged, P. (1982). A Course in Phonetics. New York: Harcourt Brace Jovanovish College Publishers.

Ladeforged, P. , & Broadbent, D. E. (1957). Information conveyed by vowels. J. Acoust. Soc. Am. 29, 98-104.

Lancker, D. V. , Kreiman, J. , & Emmorey, K. (1985). Familiar voice recognition: Patterns and parameters. Part I: Recognition of backward voices. Journal of Phonetics, 13, 19-38.

LaRiviere, C. (1977). Some acoustic and perceptual correlates of speaker identification. Proc. of 7th International Congress of Phonetic Sciences, Montreal, 558-562.

Laukkanen, A., Vilkman, E., & Oksanen, H. (1996). Physical variations related to stress and emotional state: a preliminary study. J. Phonetics, 24, 313-335.

Laver, J. (1975). Individual features in voice quality. Ph.D thesis, University of Edinburgh.

Laver, J. (1980). The Phonetic Description of Voice Quality. Cambridge: Cambridge University Press.

Lhote, E. , & Haidar, L. (1990). Speaker verification by a vocal proxemy cue. Proc. of the Tutorial and Research Workshop on Speaker Characterisation, Edinburgh, 149-154.

Li, K. P. , Dammann, J. E. & Charpman, W. D. (1966). Experimental studies in speaker verification using an adaptive system. J. Acoust. Soc. Am. 40, 966-978.

Li, K. P. , Hughes, G. W. , & House, A. S. (1970). Approaches to the characterisation of talker differences by statistical operations on speech spectra. J. Acoust. Soc. Am. 47, 66(A).

Liberman, A. M. , Delattre, P. C., Gerstman, L. J. & Cooper, F. S. (1956). Tempo of frequency change as a cue for distinguishing classes of speech sounds. Journal of Experimental Psychology 52, 127-137.

Lieberman, P. , & Blumstein, S. (1988). Speech Physiology, Speech Perception, and Acoustic Phonetics. Cambridge: Cambridge University Press.

Linde, Y. , Buzo, A. , & Gray, R. M. (1980). An algorithm for vector quantizer design. IEEE Trans. Commun. COM-28, 84-95.

Luck, J. E. (1969). Automatic speaker verification using cepstral measurements. 1969. J. Acoust. Soc. Am. 46, 1026-1032.

Lyons, J. (1977). Semantics. Cambridge: Cambridge University Press.

Macieod, I. , Chen, F. , Milliar, B. , & Laverty, W. (1994). Optimal Cohort Design in VQ-Distortion Based Text-Independent Speaker Verification. Proceedings of the Fifth Australian International Conference on Speech Science and Technology, 750-755..

Matsui, T. & Furui, S. (1991). Text-independent speaker recognition using vocal tract and pitch information. Proc. ICSLP, Kobe, 5.3.

McCarthy, J. J. (1988). Feature geometry and dependency: A review. Phonetica, 43, 84-108.

McConnnell, S. (1974). Intonation in a man's world. Paper presented at the American Anthropological Association Annual Meeting, Mexico City.

McGehee, F. (1937). The Reliability of the identification of human voice. J. Gen. Psychol. 17, 249-271.

Milenkovic, P. H. , & Read, C. (1992). Cspeech Version 4 Laboratory Automation Reference. University of Wisconsin-Madison.

Millar, B. , Chen, F. , & Wagner, M. (1994). The efficacy of cohort normalisation in a speaker verification task under different types of speech signal For Robust User-Conscious Secure Transactions. Proceedings of the Fifth Australian International Conference on Speech Science and Technology, 850-855, Perth, Australia.

Moore, G. E. (1939). Personality traits and voice quality deficiencies. Journal of Speech Disorders 4, 33-36.

Naik, J. M. (1990). Speaker Verification: A Tutorial. IEEE Communications Magazine, 42-48.

Naik, J. M. , & Doddington, G. R. (1986). High performance speaker verification using principal spectral components. Proc. ICASSP-86, 881-884.

Naik, J. M. , & Doddington, G. R. (1987). Evaluation of a high performance speaker verification system for access control, Proc. ICASSP-87, 2392-2395.

Nolan, F. (1983). The Phonetic Bases of Speaker Recognition. Cambridge: Cambridge University Press.

Nygaard, L. C., Sommers, M. S. , & Pison, D. B. (1994). Speech perception as a talker-contingent process. Psychol. Sci. 5, 42-46.

O'Connor, J. D. , Gerstman, L. J. , Liberman, A. M. , Delattre, P. C & Cooper, F. S. (1957). Acoustic cues for the perception of initial /w,y,r,l/ in English. Word 13, 24-43.

O'Shaughnessy, D. (1986). Speaker recognition. IEEE ASSP Magazine, 4-17.

O'Shaughnessy, D. (1987). Speech Communication: Human & Machine. Reading, MA: Addison-Wesley.

O'Shaughnessy, D. (1996). Critique: Speech perception: Acoustic or articulartory?. J. Acoust. Soc. Am. 99, 1726-1729.

Ohala, J. (1983). The Origin of sound patterns in vocal tract constraints. In P.F. Macneilage (Ed.), The Production of Speech, New York: Springer-Verlag.

Paliwal, K. K. (1984). Effectiveness of different vowel sounds in automatic speaker identification. Journal of Phonetics, 12, 17-21.

Palmeri, T. J. , Goldinger, S. D. , & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. J. Exp. Psychol. 19, 1-20.

Pellowe, J. , & Jones, V. (1978). On intonational variability in Tyneside speech. In P. Trudgill (Ed.), Sociolinguistic Patterns in British English, London.

Perkell, J. S. Hillman, R. E. & Holmberg, E. B. (1994). Group differences in measures of voice production and revised values of maximum airflow declination rate. J. Acoust. Soc. Am. 96, No. 2. 695-698

Peterson, G.E., & Barney, H. . (1954). Control methods used in a study of the identification of vowels. J. Acoust. Soc. Am. 24, 175-184.

Picone, J. (1990). Continuous speech recognition using Hidden Markov Models. IEEE ASSP Magazine, 26-41.

Pollack, I. , Pickett, J. M. , & Sumby, W. H. (1954). On the identification of speakers by voice. J. Acoust. Soc. Am. 26, 403-406.

Rabiner, L.R. & Schafer, R.W. (1978). Digital Processing of Speech Signals. Englewood Cliffs, NJ: Prentice Hall.

Rabiner, L. , & Juang, B-H. (1993). Fundamentals of Speech Recognition. Englewood Cliffs, NJ: Prentice Hall.

Rochet, B. , & Chen, F, (1992). Acquisition of the French VOT contrasts by adult speakers of Mandarin Chinese. Proc. ICSLP, 273-276.

Rosenberg, A. E. , & Soong, F. K. (1992). Recent research in automatic speaker recognition. In S. Furui (Ed.), Advances in Speech Signal Processing(pp. 701-738). New York: Marcel Dekker, Inc.

Rosenberg, A. E. , Lee, C-H. , & Soong, F. K. (1994). Cepstral channel normalization techniques for HMM-based speaker verification. Proc. ICSLP, 1835-1838.

Sagayama, F. S. , & Furui, S. (1991). Line spectrum pair frequency based distance measure for speech recognition. Proc. ICSLP, No. 13.1, 521 -524.

Sambur, M. R. (1976). Speaker recognition using orthogonal linear prediction", IEEE Trans. on ASSP 24, 283 - 289.

Samour, M. R. (1975). Selection of acoustic features for speaker identification, IEEE Trans. on ASSP, 23, 176-182.

Sapir, E. (1926-1927). Speech as a personality trait. American Journal of Sociology 32, 892-905.

Savic, M. , & Gupta, S. K. (1990). Variable parameter speaker verification system based on Hidden Markov Modeling. Proc. IEEE Interna. Conf. ASSP, S.5.7, 281-284.

Scherer, K. R. , & Giles, H. (1979). Social Markers in Speech. Cambridge: Cambridge University Press.

Schwartz, M. F. , & Rine, H. E. (1968). Identification of speakers from whispered vowels. J. Acoust. Soc. Am. 44, 1736-1737.

Sharkey, J. & Folkins, J. W. (1985). Variability of lip and jaw movements in children and adults: implications for the development of speech motor control. Journal of Speech and Hearing Research 28, 8-15.

Shridhar, M. , & Mohankrishnan, N. (1982). Text-independent speaker recognition: a review and some new results. Speech Communication 1, 257 - 267.

Siegman, A. W. , & Pope, B. (1965). Effects of question specificity and anxiety-producing messages on verbal fluency in the initial interview. Journal of Personality and Social Psychology 2, 522-530.

Soong, F. K. , & Rosenberg. A. E. (1988). On the use of instantaneous and transitional spectral information in speaker recognition. IEEE Trans. on ASSP 36, 871-879.

Spencer, L.E. (1988). Speech characteristics of male-to-female transsexuals: A perceptual & acoustic study. Folia Phoniatrica, 40, 31-42.

Stathopoulos, E. T. (1995). Variability revisited: an acoustic aerodynamic, and respiratory kinematic comparison of children and adults during speech. Journal of Phonetics 23, 67-80.

Stevens, K. N. (1977). Sources of inter- and intra-speaker variability in the acoustic properties of speech sounds. Proc. 7th International Congress of Phonetic Sciences, Montreal, 206-227.

Stevens, K. N. & Blumstein, S. (1981). The search for invariant acoustic correlates of phonetic features. In P. D. Eimas & J. L. Miller(eds.) Perspectives on the Study of Speech, New Jersery: Erlbaum.

Stevens, S. (1955). Measurement of Loudness. J. Acoust. Soc. Am. 27, .815-829.

Stevens. S. & Volkmann, J. (1940). The relation of pitch of frequency: A revised scale. Am. J. Psychol. 53,329-353.

Strik, H. , & Boves, L. (1992) Control of fundamental frequency, intensity and voice quality in speech. Journal of Phonetics 20, 15-25.

Su, L-S. , Li, K-P. , & Fu, K. S. (1974). Identification of speakers by use of nasal coarticulation. J. Acoust. Soc. Am. 56, 1876-1882.

Subtelny, J. , Li, W. , Whitehead, R. & Subtelny, J. D. (1989). Cephalometric and cineradiographic study of diviant resonance in hearing-impaired speakers. Journal of Speech and Hearing Disorders 54, 249-263.

Suomi, K. (1984). On talker and phoneme information conveyed by vowels: A whole spectrum approach to the normalization problem. Speech Communication 3, 199-209.

Takefuta, Y. , Jancosek, E. G. , & Brunt, M. (1971). A statistical analysis of melody curves in the intonation of American-English. Proc. of the 7th International Congress of Phonetic Sciences, The Hague.

Tohkura, Y. (1986). A weighted cepstral distance measure for speech recognition. Proc. ICASSP-86, 761 -764.

Trubetzkoy, N. S. (1969). Principles of Phonology. Berkeley and Los Angeles: University of California Press.

Trudgill, P. (1974). Sociolinguistic patterns in British English. London: Arnold.

Wagner, M. , Chen, F. , Macleod, I. , Millar, B. , Ran. , S. , Tridgell, A. , & Zhu, X. (1994). Analysis of Type-II Errors for VQ-Distortion Based Speaker Verification. Proc. of Workshop on Automatic Speaker Recognition Identification Verification, Martigny, Switzerland, 83-86.

Wagner, M. , Chen, F. , Macleod, I. , Millar, B. , Ran, S. , Tridgell, A. , & Zhu, X. (1994). Analysis of Type-II errors for VQ-distortion based speaker verification. ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, 83-86.

Warner, J. (1971). Statistical techniques for talker identification. Bell Syst. Tech. J. 50, 1427-1454.

Weismer, G. , Martin, R. , Kent, R. D. , & Kent, J. F. (1992). Formant trajectory characteristics of males with amyotrophic lateral sclerosis. J. Acoust. Soc. Am. 91, 1085-1098.

Whalen, D. H. & Levitt, A. G. (1995). The universality of intrinsic F_0 of vowels. Journal of Phonetics 23, 349-366.

Williams, C. E. , & Stevens, K. N (1972). Emotions and speech: some acoustical correlates. J. Acoust. Soc. Am. 32,1238-1250.

Witten, H. (1982). Principles of Computer Speech. London: Academic Press.

Wolf, J. J. (1972). Efficient acoustic parameters for speaker recognition. J. Acoust. Soc. Am. 51, 2044-2056.

Xu, L. & Mason, J. S. (1991). Optimization of perceptually-based spectral transforms in speaker identification. Proc. Eurospeech, 439-442.

Xu, L. , Oglesby, J. , & Mason, J. S. (1989). The optimization of perceptually-based features for speaker identification. Proc. ICASSP , 520-523.

Yorkston, K. M. , Beukelman, D. R. , & Honsinger, M. J. (1989). Perceived articulatory adequacy and velopharyngeal function in dysarthric speakers. Archives of Physical Medicine Rehabilitation 70(4), 313-317.

Zhu, X. , Gao, Y. , Ran, S. , Chen, F. , Macleod, I. , Millar, B. , & Wagner, M. (1994). Text-independent speaker recognition using VQ, mixture Gaussian VQ and ergodic HMMs. ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, 55-58.

Zhu, X. , Gao, Y. , Chen, F. , Ran. , Macleod, I. , Milliar, B. , & Wagner, M. (1994). Text-Independent Speaker Recognition Using VQ, Mixture Gaussian VQ and Ergodic HMMs, Proc. of Workshop on Automatic Speaker Recognition Identification Verification, Martigny, Switzerland, 55-58.

Zhu, X. , Millar, B. , Macleod, I. , Wagner, M. , Chen, F. & Ran, S. (1994). A Comparative Study of Mixture-Gaussian VQ, Ergodic HMMs and Left-to-Right HMMs for Speaker Recognition. ISSIPNN'94, Hong Kong, 618-621.

APPENDICES

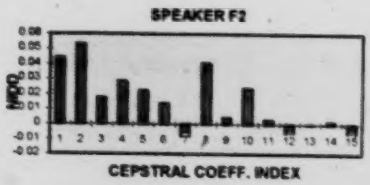
Appendix 1-A Individual Speakers' NIDD Score Distributions in the Cepstral Coefficients

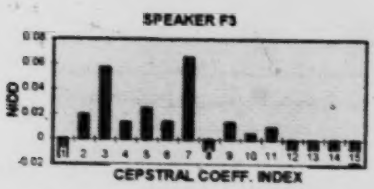
Speaker	Cepstral Coefficient														
	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15
F1	-0.05	0.01	0.02	0.01	0.04	0.03	0.01	0.02	0.03	0.00	0.00	0.01	0.00	0.01	-0.02
F2	0.04	0.05	0.02	0.03	0.02	0.01	-0.01	0.04	0.00	0.02	0.00	-0.01	0.00	0.00	-0.01
F3	-0.01	0.02	0.06	0.01	0.02	0.01	0.07	-0.01	0.01	0.00	0.01	-0.01	-0.02	-0.01	-0.02
F4	-0.01	0.05	0.07	0.04	0.01	0.02	0.00	-0.01	0.02	0.03	0.00	0.00	-0.01	-0.02	-0.03
F5	-0.02	0.02	0.04	0.02	0.00	0.01	0.01	0.01	0.01	0.01	0.00	-0.01	-0.01	-0.01	0.00
F6	-0.02	0.02	0.05	0.02	0.03	0.00	0.03	-0.02	0.02	0.02	0.00	-0.01	-0.01	0.00	-0.01
F7	-0.01	0.04	0.09	0.03	0.02	-0.01	0.01	0.01	0.01	0.00	0.00	-0.02	0.01	0.00	0.00
F8	-0.04	0.00	0.06	0.01	0.01	-0.01	0.05	0.03	0.02	0.00	0.02	-0.01	0.01	-0.01	0.00
M1	-0.09	-0.04	0.02	0.04	0.03	0.02	0.04	0.03	0.01	0.00	0.02	0.02	0.02	0.01	0.00
M2	-0.05	0.04	0.03	0.02	0.02	0.03	0.03	0.02	0.04	0.03	-0.01	-0.01	0.00	0.01	-0.02
M3	0.00	0.02	0.01	0.01	0.01	0.01	0.01	0.04	-0.01	0.04	-0.01	0.00	0.00	0.01	0.02
M4	0.02	-0.01	-0.01	-0.01	0.01	0.03	0.01	0.03	0.01	0.02	0.00	0.02	0.02	0.03	-0.02
M5	-0.04	0.05	0.03	0.03	0.01	0.00	0.02	0.01	-0.01	0.02	0.02	0.01	0.00	-0.01	0.00
M6	-0.11	0.01	0.07	0.03	0.02	0.00	0.02	0.04	0.02	0.04	-0.01	0.01	-0.03	0.01	0.00
M7	-0.03	0.02	0.01	-0.01	0.01	0.01	0.02	0.02	0.05	0.04	0.02	0.00	-0.01	-0.01	0.00
M8	-0.02	0.00	0.00	0.04	0.01	0.02	0.01	0.03	0.00	0.01	0.00	0.03	0.00	0.01	-0.02

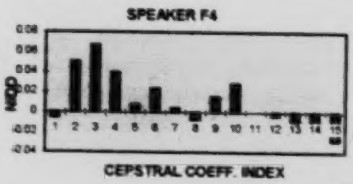
Appendix 1-B Individual Speakers' NIER Score Distributions in the Cepstral Coefficients

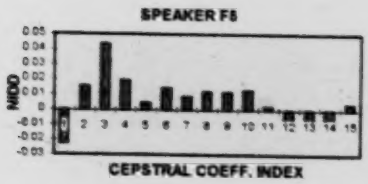
Speaker	Cepstral Coefficient														
	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15
F1	-0.04	0.03	0.09	0.04	0.04	0.03	0.01	0.00	0.09	0.02	0.01	0.01	-0.04	0.01	-0.02
F2	-0.01	-0.01	-0.01	0.00	0.01	0.01	0.00	-0.01	-0.01	0.00	0.00	0.00	0.01	-0.01	0.01
F3	0.00	0.04	0.04	0.02	-0.01	0.04	0.01	0.03	0.01	0.01	-0.01	-0.02	-0.02	-0.02	-0.01
F4	0.05	0.06	0.06	0.05	0.02	0.03	0.03	0.01	0.03	0.03	-0.02	0.03	0.00	0.01	0.01
F5	-0.01	0.04	0.03	0.06	0.03	0.01	0.05	0.02	0.06	0.01	0.02	-0.01	-0.03	-0.04	0.01
F6	0.01	0.03	0.04	0.01	0.04	0.05	0.03	0.01	-0.03	-0.01	-0.02	-0.01	-0.04	-0.01	0.02
F7	-0.01	0.01	0.04	0.03	0.03	0.01	0.04	-0.01	-0.01	0.03	-0.01	-0.03	-0.04	-0.01	-0.03
F8	0.00	-0.01	0.04	0.03	0.05	0.00	0.02	0.01	-0.01	0.01	-0.02	-0.01	-0.01	-0.01	-0.01
M1	0.01	-0.04	0.01	0.04	0.03	0.03	0.08	0.00	-0.04	0.06	0.02	0.06	0.00	0.06	-0.02
M2	-0.03	-0.02	0.01	0.03	0.02	0.02	0.01	0.01	0.03	-0.02	-0.03	-0.01	-0.02	-0.01	-0.06
M3	-0.04	0.03	0.02	-0.01	0.01	0.03	0.01	0.03	0.01	0.04	0.03	0.01	0.02	0.01	-0.01
M4	-0.01	0.01	0.04	0.03	0.02	0.01	0.02	0.03	-0.02	0.01	-0.04	-0.02	-0.02	0.01	-0.01
M5	0.02	0.03	0.04	0.00	-0.01	-0.01	-0.03	0.00	0.00	-0.01	0.02	0.03	-0.06	0.01	-0.03
M6	-0.12	-0.08	0.02	-0.01	0.04	-0.02	0.00	0.04	0.03	-0.01	0.01	0.01	-0.04	0.04	0.01
M7	0.02	0.05	0.02	-0.01	0.01	-0.01	0.02	0.03	0.02	0.01	-0.01	0.04	-0.07	-0.05	-0.03
M8	0.07	0.01	0.00	0.04	-0.02	0.00	-0.01	0.03	0.05	0.01	0.03	0.06	-0.04	0.05	0.01

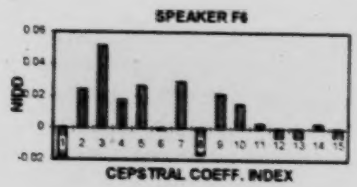
Appendix II-A Individual Speakers' NIDD Score Distributions in the Cepstral Coefficients [Female Speaker Group]

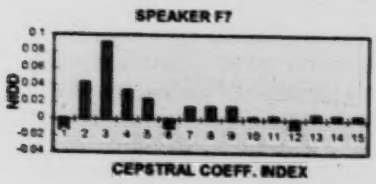


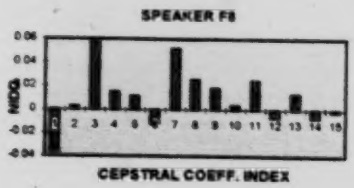




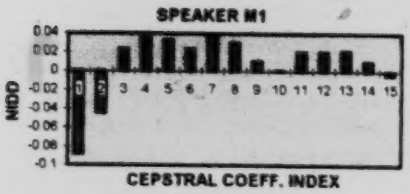


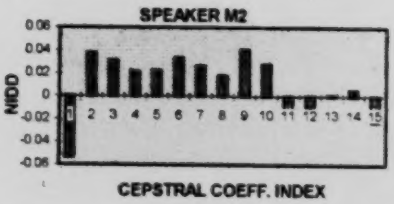


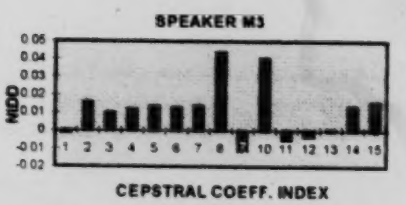


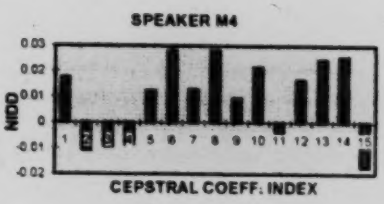


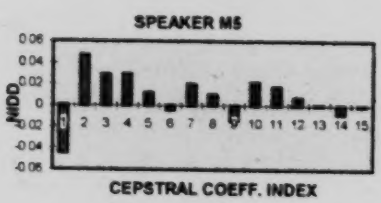
Appendix II-B Individual Speakers' NIDD Score Distributions in the Cepstral Coefficients [Male Speaker Group]

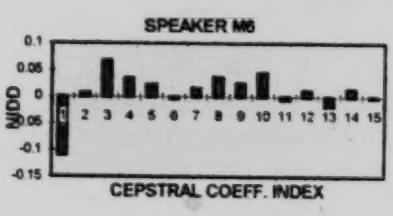


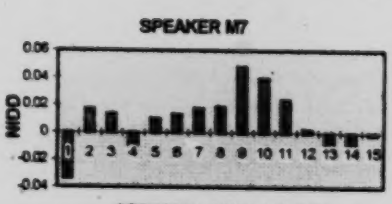


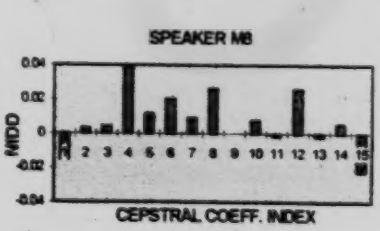




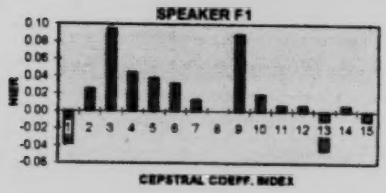


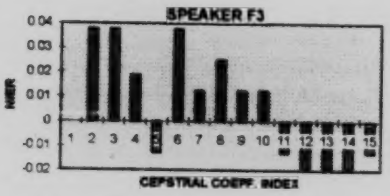


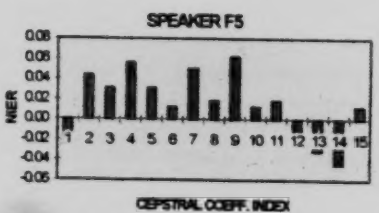


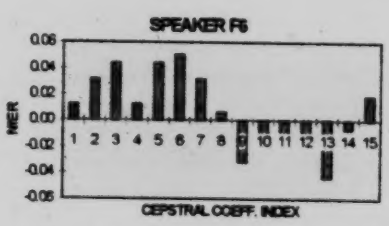


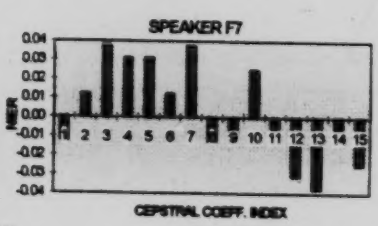
Appendix III-A Individual Speakers' NIER Score Distributions in the Cepstral Coefficients [Female Speaker Group]

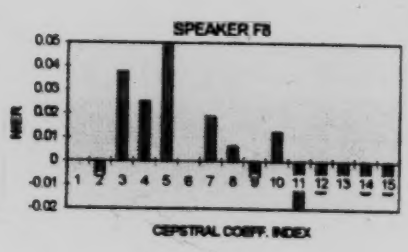




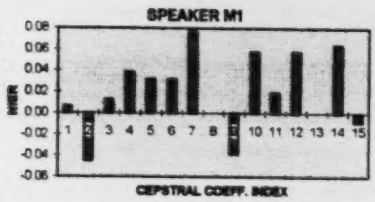


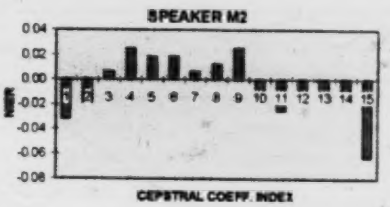


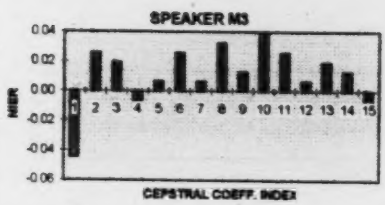


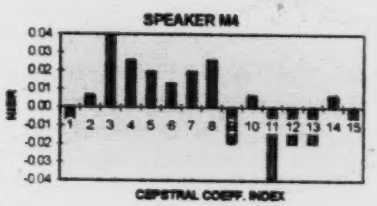


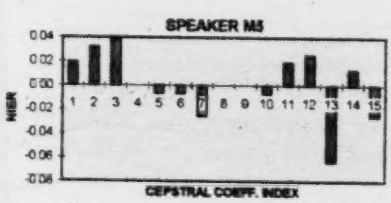
Appendix III-B Individual Speakers' NIER Score Distributions in the Cepstral Coefficients [Male Speaker Group]

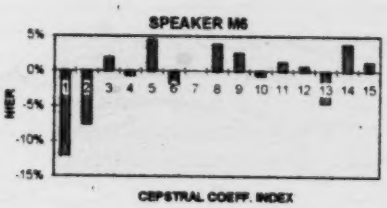


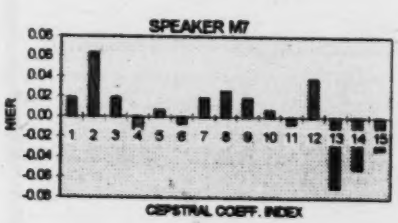


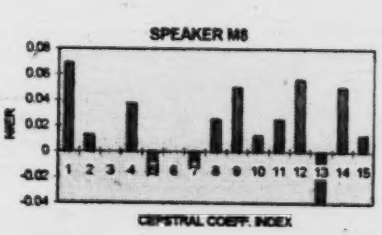








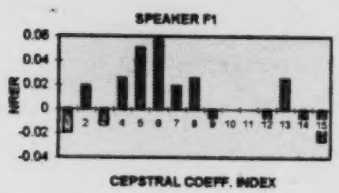


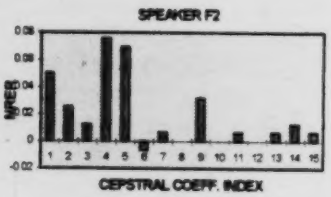


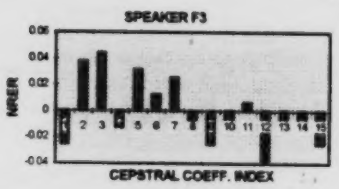
Appendix IV Individual Speakers' NRER Score Distributions in the Cepstral Coefficients

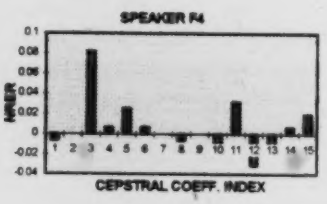
Speaker	Cepstral Coefficient														
	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15
F1	-0.02	0.02	-0.01	0.03	0.05	0.06	0.02	0.03	-0.01	0.00	0.00	-0.01	0.03	-0.01	-0.03
F2	0.05	0.03	0.01	0.08	0.07	-0.01	0.01	0.00	0.03	0.00	0.01	0.00	0.01	0.01	0.01
F3	-0.03	0.04	0.04	-0.01	0.03	0.01	0.03	-0.01	-0.03	-0.01	0.01	-0.04	-0.01	-0.01	-0.03
F4	-0.01	0.00	0.08	0.01	0.03	0.01	0.00	-0.01	0.00	-0.01	0.03	-0.03	-0.01	0.01	0.02
F5	-0.03	0.03	0.04	0.03	0.03	0.01	0.01	0.02	0.04	-0.01	0.01	-0.02	0.03	-0.03	0.01
F6	-0.01	0.03	0.08	0.08	0.02	0.03	0.04	0.01	0.01	0.00	0.01	-0.01	-0.01	0.00	-0.03
F7	0.00	0.01	0.01	0.03	0.05	0.00	0.01	0.00	0.01	0.00	0.01	0.02	0.03	0.01	0.03
F8	-0.03	0.00	0.04	0.03	0.03	0.01	0.02	-0.01	0.00	-0.01	0.00	0.00	0.00	0.00	0.03
M1	-0.08	0.02	0.04	0.01	0.03	0.01	0.10	0.01	-0.02	-0.01	-0.01	0.03	0.01	0.01	0.03
M2	-0.04	0.00	-0.01	-0.01	-0.01	0.03	-0.02	-0.03	-0.03	0.00	-0.03	0.00	-0.02	-0.03	-0.02
M3	-0.03	-0.01	0.01	0.00	0.01	0.00	0.00	0.01	0.00	0.01	-0.01	-0.03	-0.01	0.01	-0.01
M4	-0.01	-0.01	0.00	-0.01	0.08	0.01	0.01	-0.01	0.02	-0.02	0.01	0.01	0.01	0.03	0.00
M5	-0.02	0.00	0.03	0.00	0.03	0.02	0.02	0.00	0.02	0.01	-0.01	0.01	0.00	0.01	0.02
M6	-0.06	0.01	0.01	-0.01	0.02	0.03	-0.02	0.03	-0.01	0.05	-0.02	0.01	-0.03	0.01	-0.06
M7	-0.07	0.00	-0.02	0.01	0.02	-0.01	0.01	0.01	0.03	-0.03	0.01	-0.01	-0.01	-0.01	0.01
M8	0.01	0.03	0.04	0.02	0.03	0.01	-0.01	0.03	0.01	-0.02	-0.01	0.04	-0.01	0.00	0.01

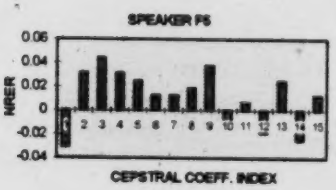
Appendix V-A Individual Speaker's NRER Score Distribution in the Cepstral Coefficients [Female Speaker Group]

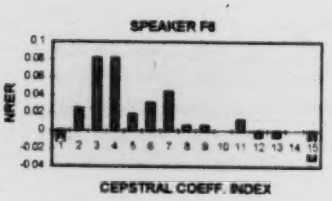


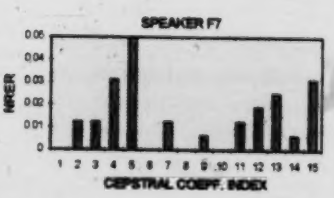


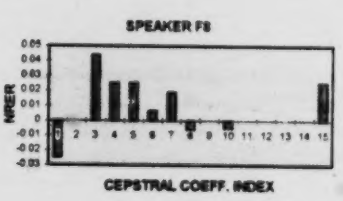












Appendix V-B Individual Speaker's NRER Score Distribution in the Cepstral Coefficients [Male Speaker Group]